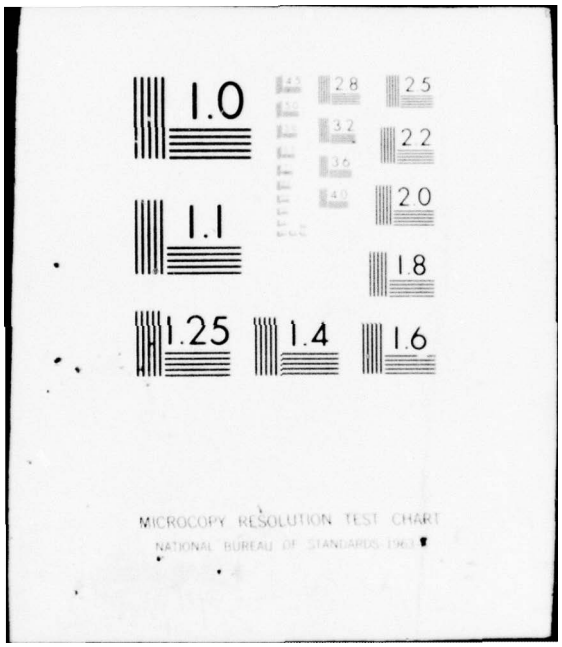
AD-A033 343 STANFORD UNIV CALIF DEPT OF OPERATIONS RESEARCH F/G 12/2
A STOCHASTIC CAPACITY EXPANSION MODEL: MODULAR TEMPORARY FACILI--ETC(U)
OCT 76 R S SHIPLEY N00014-75-C-0561
UNCLASSIFIED TR-179 NL

1 OF 1
AD
A033343

END

DATE
FILMED
2-77


1.0
1.1
1.25
1.4
1.6
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963

(12)

A STOCHASTIC CAPACITY EXPANSION MODEL: MODULAR TEMPORARY FACILITIES

by

R. SCOTT SHIPLEY

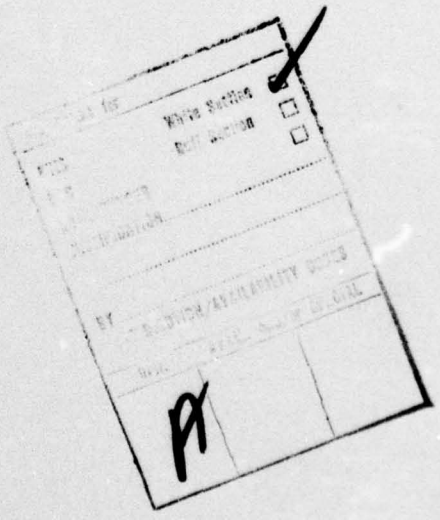
TECHNICAL REPORT NO. 179
October 4, 1976

SUPPORTED BY THE ARMY AND NAVY
UNDER CONTRACT N00014-75-C-0561 (NR-042-002)
WITH THE OFFICE OF NAVAL RESEARCH

Gerald J. Lieberman, Project Director

DEPARTMENT OF OPERATIONS RESEARCH
AND
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

TABLE OF CONTENTS



CHAPTER 1

INTRODUCTION

1.1. Preview

This study considers optimal decision strategies with regard to capacity expansion in a stochastic demand environment. Chapter 1 is an introduction to the topic at hand and provides the preliminaries upon which later model construction is based. As indicated in the next section, there are two types of facilities which must be considered in a capacity expansion model: permanent and temporary. With regard to temporary facilities, a further classification is necessary according to whether or not the temporary facilities are modular.

Model I, presented in [7], treats the case where temporary facilities are non-modular. Model II, presented herein, treats the case where temporary facilities are modular. For both cases, it is shown that the determination of optimal expansion sizes can be expressed as a single-variable, nearly-unconstrained, minimization problem of a very particular form. The solution to this problem is treated in [8], where a revised model is also introduced to demonstrate a number of generalizations that are possible in both Models I and II.

1.2. Permanent and Temporary Facilities

When one considers the growth in capacity of a singular system, such as a plant, pipeline, superhighway or school, a common phenomena

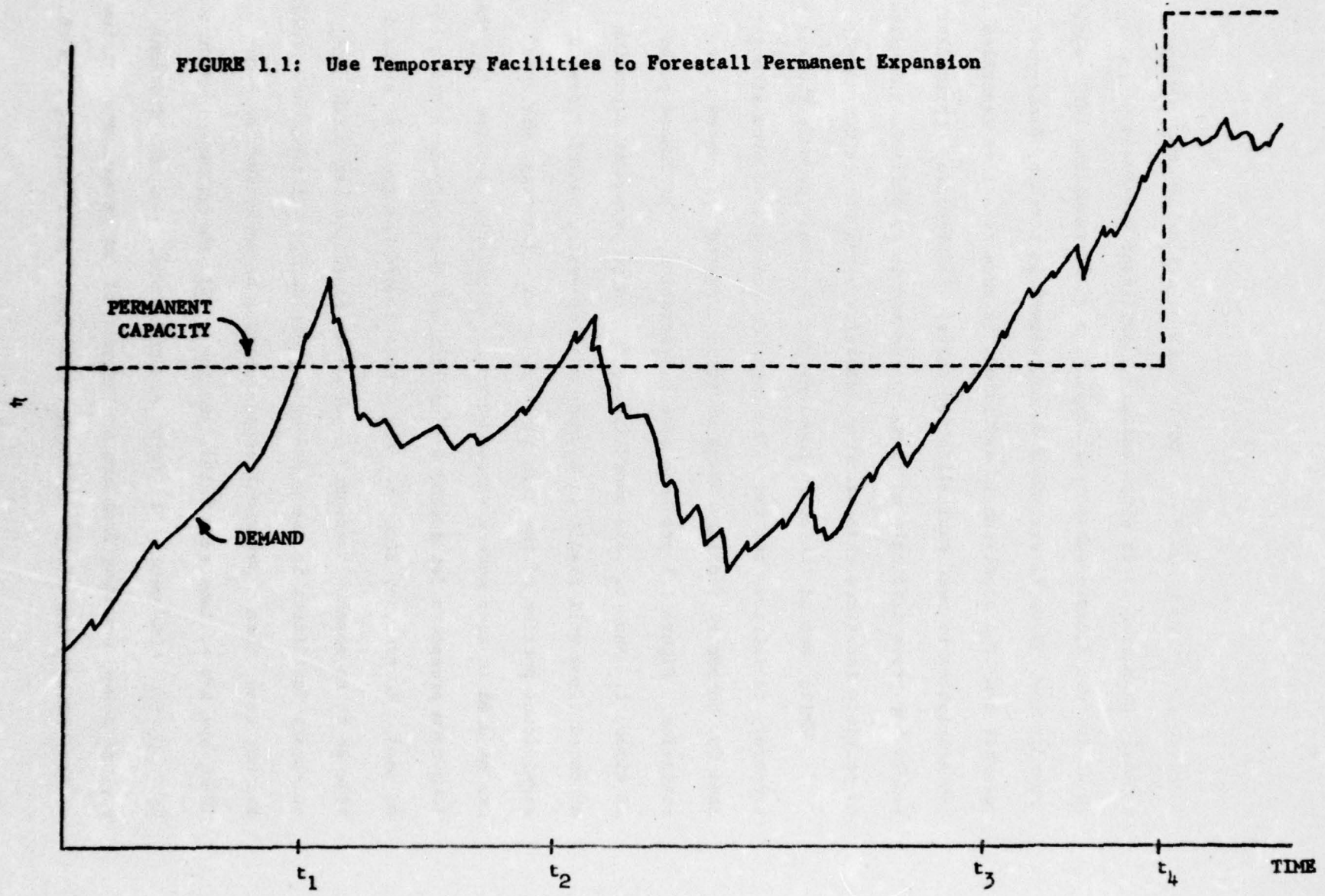
often arises. Specifically, when demand first exceeds the initial capacity of the system, a permanent capacity expansion is not immediately undertaken. Instead, temporary operating measures are instigated in order to satisfy excess demand over the short-run. In some cases, the temporary measure taken may be simply to backlog excess demand. In other cases, where backlogging is undesirable, the present system may be overloaded beyond its planned capacity in order to accommodate excess demand. In still other cases, where backlogging is undesirable and overloading is infeasible, temporary capacity may be rented from outside sources. In order to distinguish between the normal accommodation of demand and the emergency accommodation of excess demand through temporary measures, facilities will be categorized here as either permanent or temporary. That is, "temporary facilities" will denote the measures utilized to accommodate excess demand over the short-run prior to the implementation of a full-fledged expansion of permanent facilities.

The use of temporary facilities usually involves some sort of cost penalty over that of permanent facilities; otherwise, one might well ask which facilities are really temporary. Thus, the economic advantages of temporary facility usage might, on the surface, seem questionable. There are, however, two principal factors which promote the desirability of temporary facility usage. Firstly, there is the problem of uncertainty: when demand first exceeds the permanent capacity, is the excess a short-term peak or a long term trend? If the time interval of excess is to be of small duration, then an expansion of permanent facilities can be disastrous, since overhead costs on unused capacity can easily lead to financial

insolvency. Secondly, there is the problem of capitalization: permanent capacity expansion costs often exhibit significant economies of scale, which in turn dictate substantial expansion sizes requiring large capital expenditures. Even if sustained demand growth is certain, bankruptcy is possible over the short-run if sufficient revenue cannot be generated from new facilities to meet capitalization costs. Furthermore, capitalization itself may prove difficult until excess demand is of sufficient magnitude to convince investors that permanent capacity expansion is warranted.

While each of the above problems can in itself promote the use of temporary facilities, the two taken together generate an even stronger case for temporary facility usage under an expected discounted cost criterion. Figure 1.1 provides some illustration. The demand peaks at times t_1 and t_2 are short-lived; any large permanent expansion at these times will result in a great deal of wasted capacity over a significant portion of the time interval shown. Temporary measures can be used to accommodate these excesses. Similarly, at time t_3, the long-term prospects for demand are unclear and temporary facilities can be used. However, by time t_4, the excess demand has grown to significant size so as to somehow "warrant" a permanent expansion (at which time, the temporary facilities in use would be discontinued). If temporary facilities are not used, then a permanent expansion must be undertaken at time t_1. Thus, the use of temporary facilities forestalls the permanent expansion for (t_4-t_1) time units. If costs are discounted, then the discount savings alone, accruing from the prolongment of the permanent expansion capital outlay, can easily exceed any cost penalties resulting from the

FIGURE 1.1: Use Temporary Facilities to Forestall Permanent Expansion



use of temporary facilities at times t_1, t_2 and t_3.

In order to make these ideas more precise, let k denote the spares level: the difference between present permanent capacity and demand. Then -k denotes the excess demand whenever demand exceeds permanent capacity. When k is nonnegative, the permanent facilities suffice to serve all demand. However, when k is negative, excess demand exists and temporary facilities must be used. Let K denote the limit on temporary facility usage. Whenever $k = -K$ (that is, excess demand equals K) and a new unit of demand growth occurs, a permanent expansion of size $X+1$ $(X \geq K)$ is required. When permanent expansion occurs, the temporary facilities in use are discontinued and the spares level k increases to $X-K \geq 0$. If demand consists of deterministic constant growth, then the use of this type of recurrent (X,K) expansion policy results in a saw-tooth pattern for the spares level, as illustrated in Figure 1.2.

The value K serves as a "trigger", forcing a permanent expansion whenever a new unit of demand growth occurs while $k = -K$. Thus, $K+1$ represents the point at which excess demand warrants a permanent expansion. In cases of overloading, the value K may be determined based on physical constraints governing the degree of overloading that can safely be incurred. In other cases, K may be determined according to a judgment concerning the willingness of investors to participate in capitalization and the prospects of generating sufficient revenue to meet capitalization costs, given that excess demand has reached the value K. In such cases where K is predetermined, the parameter of interest will be $X^*(K)+1$, the

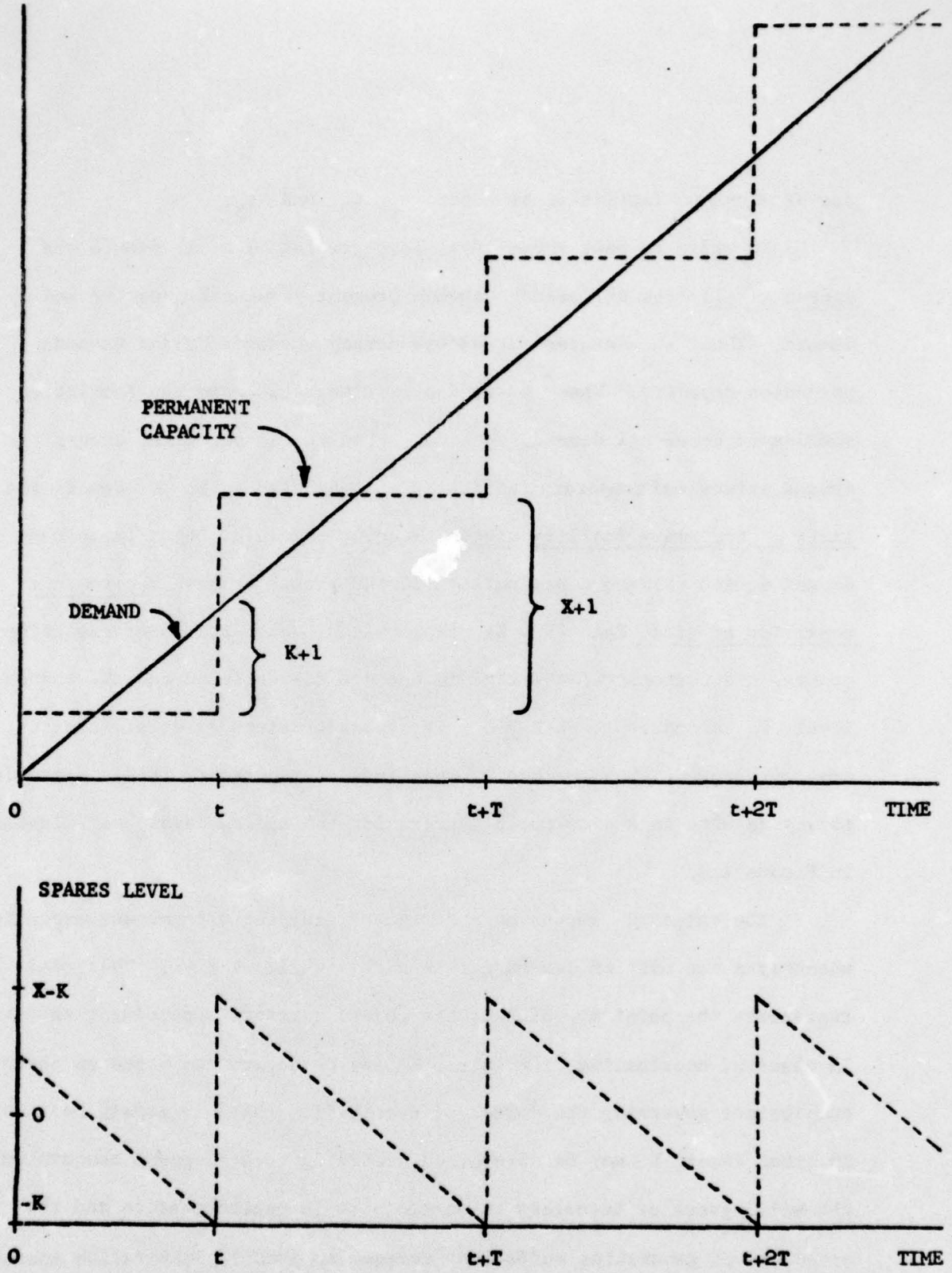


FIGURE 1.2: Recurrent (X,K) Expansion Policy for Linear Demand; Spares Level = Permanent Capacity-Demand.

optimal permanent expansion size to be utilized in conjunction with the temporary facilities usage limit K. In other cases, a value K* will be determined based on a cost minimization over all feasible operating policies (X,K). Given a procedure for determining X*(K), this last problem reduces to that of a cost minimization over all feasible operating policies (X*(K), K).

1.3. Modular Temporary Facilities

Modular temporary facilities can be viewed as a specific case of jobletting where temporary facilities can be rented from outside sources in a fixed increment size.* An instance of this type is the use of mobile "barracks" facilities (e.g., house trailers) as auxiliary classrooms to alleviate conjestion in overcrowded schools.

Another instance of this type is the use of "pair-gain" devices to augment the capacity of over-subscribed telephone cables. To illustrate, consider Figure 1.3. In order to simplify the illustration, consider only the one-way communication from a subscriber telephone to the switching equipment at the subscriber's local telephone office. As shown in Figure 1.3a, this communication is normally implemented by the dedication of a single wire-strand from the subscriber telephone to the office. This single wire is typically one strand of a telephone cable between the subscriber's neighborhood and the office. Suppose that the demand for

*Note: Modular cost behavior can also be found in cases of backlogging and overloading; however, it is more prevelant in jobletting or rental situations.

TELEPHONE PAIR-GAIN SYSTEMS

FIGURE 1.3a: Wire-Only

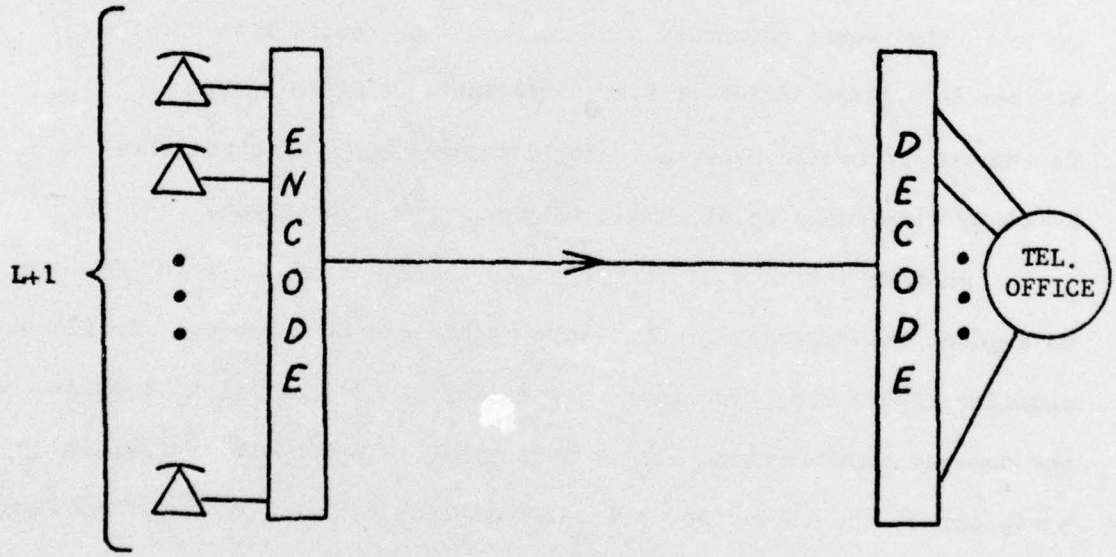


FIGURE 1.3b: Pair-Gain L:1

telephones in this neighborhood increases to a point at which all wire-strands of the existing cable are used. The telephone company, being required to provide service to all soliciting subscribers, must somehow increase the cable capacity in order to satisfy new demands. One obvious possibility is the immediate placement of a new cable (i.e., an immediate permanent expansion). Another possibility is the implementation of multiple-party service; however, this alternative is usually unexceptable to consumers and is, in fact, not used in many areas. Another alternative to the placement of an additional cable is the utilization of pair-gain devices on the existing cable. As illustrated in Figure 1.3b, a pair-gain device consists of electronics that permit the transmission of a number of simultaneous conversations over a single wire-strand by encoding the conversations at one end and decoding the conversations at the other end. As shown in the figure, these devices are typically available in a fixed modular size allowing the transmission of L additional conversations over a single wire-strand; the "pair-gain ratio" of each such module is said to be "L to 1".

The cost characteristic which distinguishes modular temporary facilities from non-modular temporary facilities is the existence of instaneous charges over and above the normal rental costs. Specifically, an installation charge is typically incurred whenever an additional module is engaged and similarly, a removal charge is incurred whenever a module is returned to the outside supplier. Thus, in the case of modular temporary facilities, there is a sub-optimization problem with regard to how modules should be engaged and returned in order to avoid excessive installation and removal charges.

1.4. The Poisson Demand Model

In order to account for the promotion of temporary facility usage due to uncertainty, demand will be recognized as being stochastic in nature. Specifically, it will be assumed that demand increases ("arrivals") are characterized by a Poisson process [6] at rate $\lambda_1 > 0$. Similarly, it will be assumed that demand decreases ("departures") are characterized by a Poisson process at rate $\lambda_2 > 0$. Furthermore, the arrival and departure processes will be presumed to be independent of each other and also independent of the expansion policy chosen.

Given the above demand characterization, define an "event" as either an arrival or a departure. Then an equivalent, and more convenient, characterization is as follows: events occur according to a Poisson process at rate $\lambda = \lambda_1 + \lambda_2$; the probability that an event is an arrival equals $p = \lambda_1/\lambda$; and the probability that an event is a departure equals $q = \lambda_2/\lambda$ ($q = 1-p$) [6]. That is, the demand process treated here can be viewed as a random walk with exponential inter-event times.*

It should be noted that the above demand characterization is pointed toward systems providing a homogeneous continuous service to a fairly stable clientele, such as a power plant or a telephone service area. It is really not intended to model actual clientele movement for cases where customers are rapidly flowing in and out, such as a job shop. In these cases, it is suggested that "customer demand" be viewed as the maximum usage level required of facilities during short intervals of time (e.g., weeks or months).

---

*Notice that the departure process is untruncated: even when the demand level is zero, departures are allowed. The viability of this assumption is demonstrated in Section 1.7 of [7].

It should also be noted that the demand process is assumed time-stationary. For this reason, the system being modelled must be in the midst of a relatively stable period of either sustained growth or sustained negative growth.

Finally, notice that the arrival and departure rates are presumed constant with regard to the demand level. From a practical point of view, this is convenient since only two estimates -- a single arrival rate and a single departure rate -- are necessary to implement the model. However, in most realistic situations, it appears that the arrival and departure rates will vary according to the system demand level. In recognition of this fact, the revised model of [8] permits these rates to vary.

1.5. Costs

The optimization criterion used here will be the minimization of all future expected discounted costs; r will denote the continuous discount rate, $0 < r < 1$. Three types of costs will be allowed: permanent expansion costs, permanent facility operating charges and temporary facility charges. All costs are assumed to be time-stationary.

Permanent expansion costs will be denoted by $g(\cdot)$, where $g(X)$ is the cost of a permanent expansion of size $X+1$. Notice that $g(\cdot)$ is presumed to be a function only of the expansion size, independent of the level of existing permanent capacity and the value of the temporary facilities usage limit, K. This restriction is relaxed in the revised model of [8].

It is important to note that $g(0)$ represents the cost of a unit expansion. Thus, $g(0)$ is comprised of the fixed (i.e., "setup") expansion cost plus the marginal cost for the first unit of increased capacity. Hence, any presumption regarding a specific functional form (e.g., convexity) for g over $[0,\infty)$ will not preclude the existence of a fixed expansion cost.

With regard to other costs, the temporary facility charges are of principal interest in this study. In order to provide sufficient generality with regard to these costs, the permanent facility operating charges will be presumed to be proportional to the amount of permanent capacity used. The following theorem indicates the modelling simplifications that result from this assumption.

Theorem 1.1. Suppose that permanent facility operating charges are proportional to the amount of permanent capacity used, at rate γ per unit-time. Then an equivalent cost model is given as follows:

(i) When permanent facilities satisfy all demand, no operating charges are incurred.

(ii) When temporary facilities are necessary (i.e., $k < 0$) marginal costs equal to the temporary facility charges, less an adjustment $\gamma|k| = -\gamma k \geq 0$ per unit-time, are incurred.

Proof. See Theorem 1.1 of [7].

The presumed independence of the permanent operating costs from the permanent facility capacity may not be that restrictive, since proportional fixed operating charges which depend solely on the permanent capacity level can be included in the expansion cost function g. To illustrate, suppose that the actual permanent operating costs include a proportional fixed charge of ψ per unit-time for each unit of permanent capacity available. Let Y_0 denote the initial permanent capacity. The expected discounted cost resulting from the ψ charge on the initial capacity is then $\int_0^\infty Y_0 \psi\, e^{-rt}\, dt = \psi r^{-1} Y_0$; this cost is independent of the expansion policy chosen. The expected discounted costs resulting from the ψ charge on future capacity expansion increments can then be accounted for by defining g as

$$g(X) = g_e(X) + \int_0^\infty (X+1)\psi\, e^{-rt}\, dt$$

$$= g_e(X) + \psi r^{-1}(X+1) ,$$

where g_e represents the actual expansion function alone.

Temporary facility costs will be represented by a per-module rental charge π per unit-time, installation charge c, and removal charge d; notice that the later two costs are instantaneous charges.

All restrictions (except time-stationarity) on the facility charges, both permanent and temporary, are also relaxed in the revised model of [8].

1.6. Remarks and Notation

Related results are discussed at length in Section 1.6 of [7]. Pertinent related models include those of Manne [5], Freidenfelds [2], and Koontz [4]. [7] also demonstrates a near-analogy between the model at hand and single-item inventory models with both returns and backlogs.

As introduced previously in this chapter, $\mathfrak{D}(t)$ will denote the demand at time $t \geq 0$, λ will denote the event rate with arrival probability p and departure probability q, k will denote the spares level, K will denote the temporary facilities usage limit, $X+1$ will denote the expansion size and g will denote the permanent expansion cost function, where $g(X)$ represents the cost for an expansion of size $X+1$.

$\mathbb{R}^m$ will denote the space of all real vectors having m components, $m \geq 1$. Both row and column vectors will be utilized; however, a vector is a column vector unless otherwise noted. Given $V \in \mathbb{R}^m$, V^T will denote the transposed vector. $R^{m \times n}$ will denote the space of all real matrices having m rows and n columns. Given $A \in \mathbb{R}^{m \times n}$, $A_{i.}$ will denote row i of A and similarly, $A_{.j}$ will denote column j of A ($A_{i.}^T \in \mathbb{R}^n$ and $A_{.j} \in \mathbb{R}^m$).

CHAPTER 2

MODEL II: THE MODULAR CASE

2.1. Introduction

This chapter characterizes the untruncated Poisson model under two basic assumptions:

Assumption 2.1. Permanent facility operating charges are proportional to the amount of permanent capacity used, at rate γ_1 per unit-time.

Assumption 2.2. Modular temporary facilities are available to satisfy excess demand. Each module provides L units of capacity. The installation of a module incurs an instantaneous connect charge $c > 0$; similarly, the disconnection of a module incurs an instantaneous charge $d > 0$. While a module is connected, a rental charge of $\pi > \gamma_1 L$ per unit-time is also incurred.

Assumption 2.1 is a restatement of the hypothesis for Theorem 1.1. Therefore, using Theorem 1.1, the construction of Model II will proceed under the equivalent assumption:

Assumption 2.1'.

(i) When permanent facilities satisfy all demand, no operating costs are incurred.

(ii) When temporary facilities are necessary (i.e., $k < 0$) marginal costs equal to the temporary facility charges, less an adjustment $\gamma_1|k| = -\gamma_1 k \geq 0$ per unit-time, are incurred.

In [7] (Example 2.5), it was shown that if the temporary facility charges satisfy Assumption 2.2, and if the utilization of temporary modules is required to track the demand pattern, then Model I of [7] was applicable. This situation can be represented in a transition diagram, as shown in Figure 2.1. In the diagram, spares levels are depicted vertically and spares levels using the same number of temporary modules are aligned in columns. Each circled number denotes the number of temporary modules necessary for the corresponding spares level given horizontal to the circle in the righthand margin. Transitions occurring due to departures are denoted by solid lines, while transitions occurring due to arrivals are denoted by dashed lines.

Connection and disconnection (i.e., installation and removal) charges are denoted by triangles. Whenever all modules are fully used ($k = -nL$ for some integer n) and an arrival next occurs, two possibilities occur. If $k = -K$ ($= -NL$ for some integer N), then a permanent capacity expansion is undertaken and the N modules are disposed of; otherwise ($n < N$), a new module is connected at cost c, increasing the total number of modules in use to $n+1$. Similarly, whenever a module is being used to serve a single customer (i.e., $k = -(n-1)L-1$ for some $n \geq 1$) and a departure next occurs, a module is disconnected at cost d, decreasing the number of modules in use to $n-1$. In addition to the instantaneous connect and disconnect charges, a cost rate of π per unit-time is incurred for each module in use.

Given "free rein" on module usage, optimal disconnections will usually not track the demand pattern. Specifically, the optimization

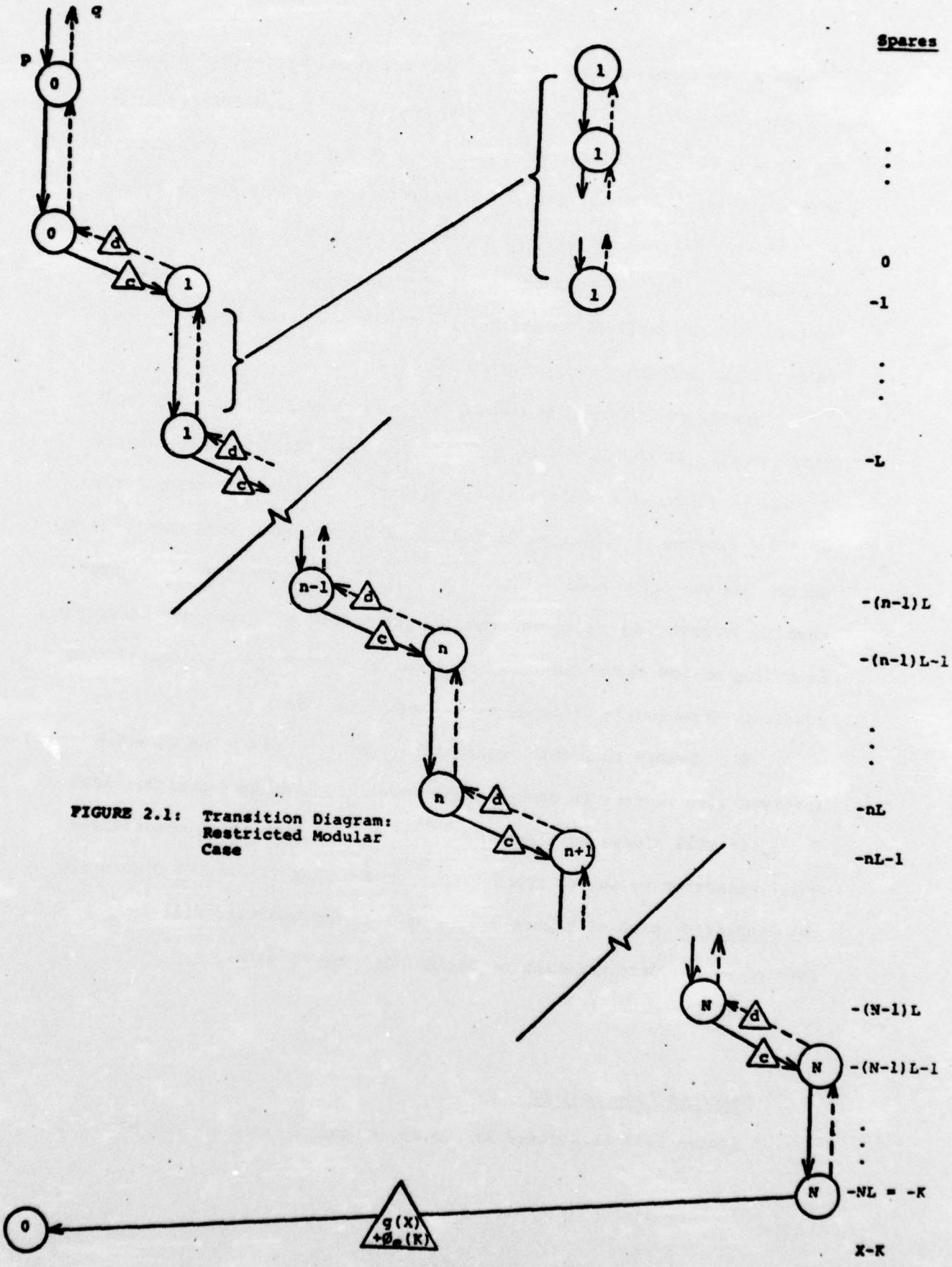


FIGURE 2.1: Transition Diagram: Restricted Modular Case

dilemma to be resolved is whether or not one should forestall a feasible disconnection, willfully incurring the penalty of rental charges on an unnecessary module, to protect against a subsequent near-future reconnection. To illustrate, consider Figure 2.2. Suppose the current spares level is $k = -1$; that is, one unit of temporary capacity in a single module is necessary. Suppose that demand now decreases by one (i.e., a departure), so that the spares level passes to $k = 0$; that is, the temporary module is no longer necessary to serve demand.

Question: should the module, now unnecessary, be disconnected immediately? If the disconnection is made, then the disconnect charge d will be incurred and there is the risk of a near-future reconnection, at added expense c, whenever the net demand (from now) next increases by unity. On the other hand, if the disconnection opportunity is declined, then the rental charges on the unnecessary module will continue to accrue. Expanding on the above question, interest will be focused on determining precisely when module disconnections should be made.

With regard to module connections, there is really no question involved with respect to economic optimization. Assuming neglible lead times, it will always be economically advantageous to defer connections until necessary to serve demand (since $c > 0$ and $\pi > \gamma_1 L \geq 0$). Hence, the connection path of Figure 2.1 is optimal and analysis will now be focused on the determination of optimal disconnect paths.

2.2. Disconnection Policies

States will be denoted by (n,k), where n denotes the number

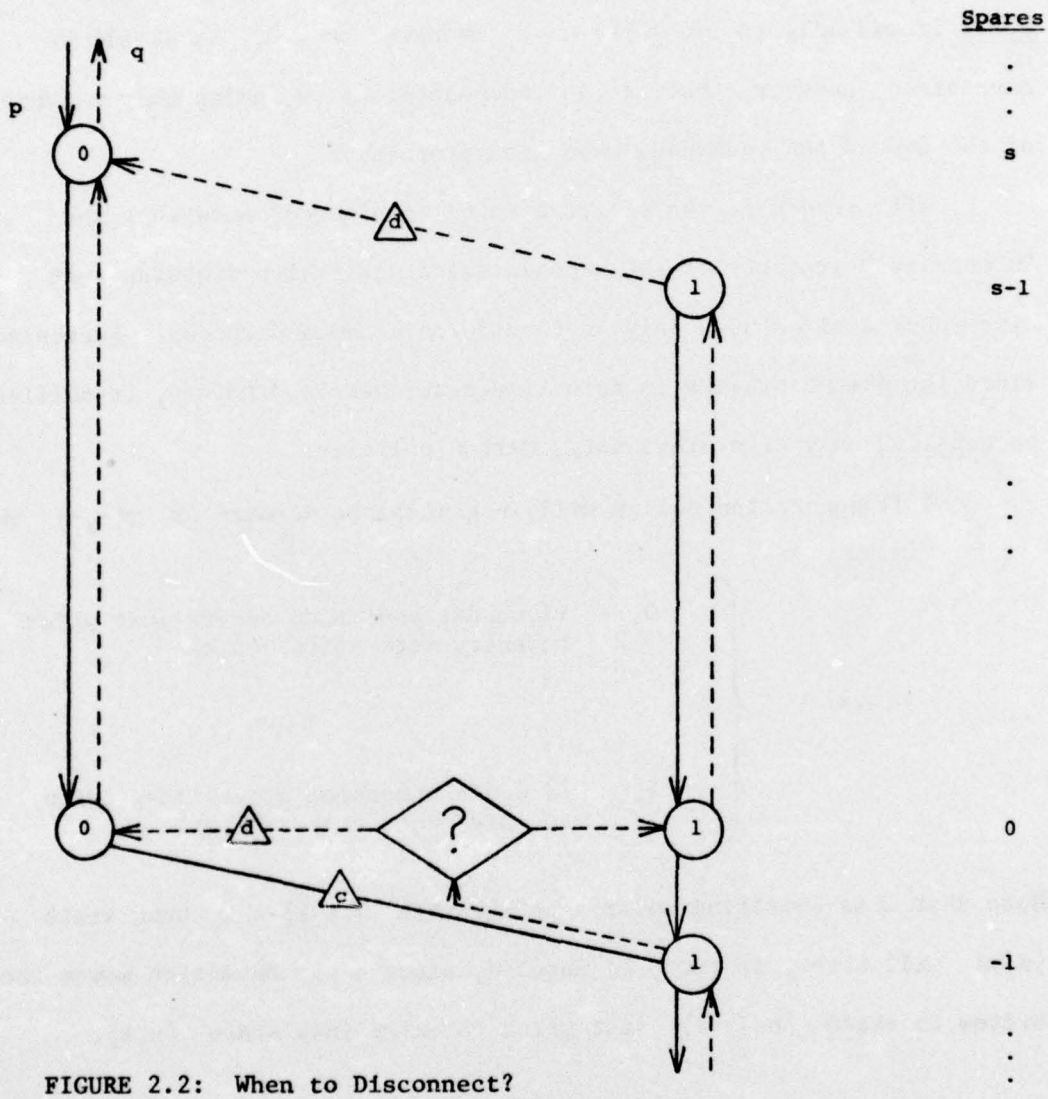


FIGURE 2.2: When to Disconnect?

of temporary modules currently connected and k denotes the current spares level. Since all demand must be served, (n,k) is a *feasible state* if and only if $nL \geq |k| = -k$, whenever $k < 0$. It should be emphasized, however, that k is independent of n, being only a function of the demand and permanent expansion processes.

With regard to the policies to be considered, note that the "memoryless" property of the exponential distribution dictates that disconnect actions need only be considered at demand epochs. Furthermore, since the demand process is both time-stationary and Markov, it suffices to consider only time-stationary, Markov policies.

A disconnection policy will in general be denoted by $\omega(\cdot,\cdot)$ where

$$\omega(n,k) = \begin{cases} 0, & \text{if no disconnection occurs just prior to entry into state } (n,k) \\ 1, & \text{if a disconnection occurs just prior to entry into state } (n,k). \end{cases}$$

Note that when operating under a policy with $\omega(n,k) = 1$, then state (n,k) will never, in fact, be entered, since a disconnection moves the system to state $(n-1, k)$ just prior to entry into state (n,k).

Definition 2.1. A policy $\omega(\cdot,\cdot)$ is *feasible* if and only if $\omega(0,k) = 0$ for all k and $(n-1)L \geq -k$ whenever $\omega(n,k) = 1$, $k < 0$. A feasible policy $\omega(\cdot,\cdot)$ is an *s-policy* if $\omega(\cdot,\cdot)$ is determined by nonnegative scalars $\{s_1, s_2, \ldots\}$ such that

$$\omega(n,k) = \begin{cases} 0 , & k < -(n-1)L + s_n , \\ & \qquad n \geq 1 . \\ 1 , & k \geq -(n-1)L + s_n , \end{cases}$$

An s-policy is a constant s-policy if there exists a nonnegative scalar s such that $s_n = s$, $n \geq 1$.

Notice that Figure 2.1 illustrates a constant s-policy with parameter $s = 0$; hence, this was the form of the disconnection policy considered in the restricted modular case of Model I. Figure 2.3 illustrates the form of the transition diagram when operating under a constant s-policy, assuming $0 < s < L$. Figure 2.4 illustrates the form of a constant s-policy, assuming $L < s < 2L$.

Physically, a general s-policy can be interpreted as following the rule: "whenever n modules are connected, disconnect a module if and only if atleast s_n units of unused capacity will remain available." In the next section, it is shown that a constant s-policy is an optimal disconnection policy.

2.3. The Optimality of Constant s-Policies

Definition 2.2. Let the random variable T_i denote the first time that the net demand reaches i:

$$T_i = \min\{t \geq 0 : \mathcal{D}(t) - \mathcal{D}(0) = i\}, \quad i = 0, \pm 1, \pm 2, \ldots .$$

In the absence of intervening permanent expansion, the excess demand

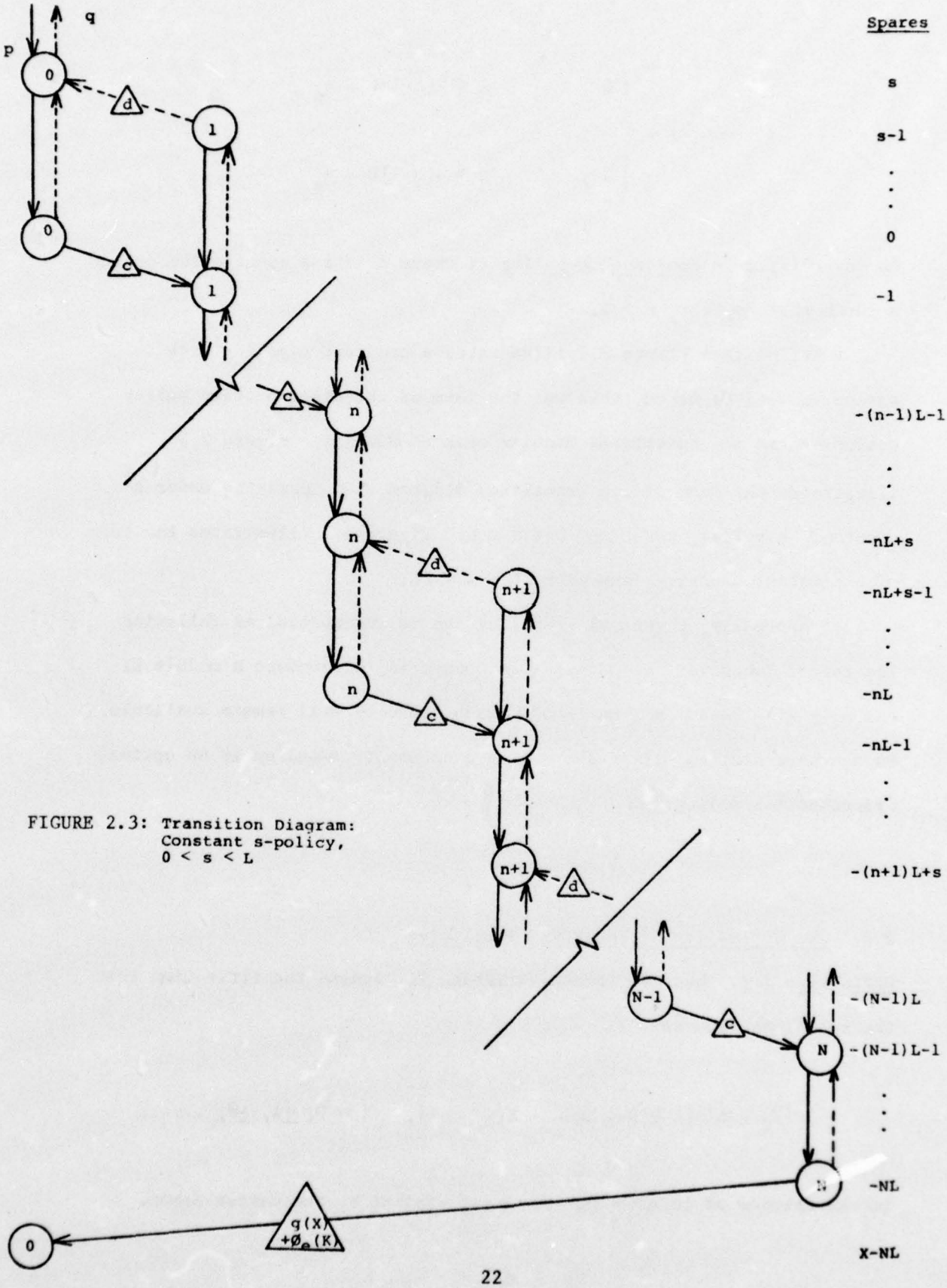


FIGURE 2.3: Transition Diagram:
Constant s-policy,
$0 < s < L$

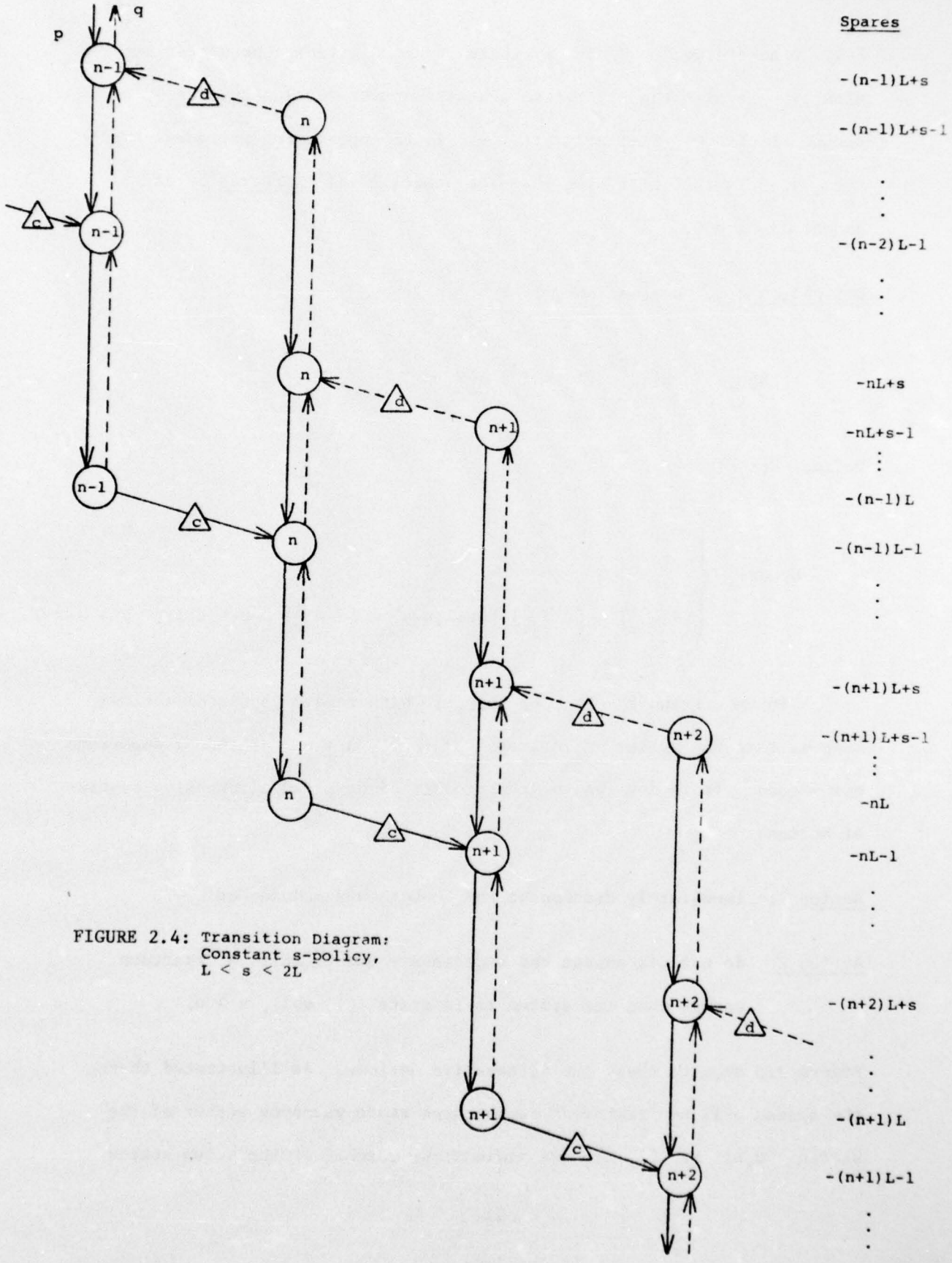


FIGURE 2.4: Transition Diagram: Constant s-policy, $L < s < 2L$

(-k) tracks demand. Hence, if there is no intervening permanent expansion, T_i is also the first time that the spares level decreases by an amount i, $i > 0$. Similarly, if there is no intervening permanent expansion, T_{-i} is the first time that the spares level increases by an amount i, $i > 0$.

Definition 2.3. Define $h(\cdot,\cdot)$ by

$$h(u,s) = E[e^{-rT_u} | T_u < T_s] \, P\{T_u < T_s\} \quad .$$

Define $H(\cdot,\cdot)$ by

$$H(u,s) = \begin{cases} 0, & u = s \\ (d - \frac{\pi}{r}) + (c + \frac{\pi}{r}) \, h(u+1,u-s) - (d - \frac{\pi}{r}) \, h(u-s,u+1), & s > u \geq 0 \end{cases}$$

To understand the role of $H(\cdot,\cdot)$ with regard to disconnections, suppose that the system is in state $(1, u-1)$, $u \geq 0$, and that a departure now occurs. It is desired to economically compare two alternative courses of action:

Action 1: immediately disconnect the unnecessary module; or

Action 2: do not disconnect the unnecessary module until a departure occurs when the system is in state $(1, s-1)$, $s > u$.

Figure 2.5 depicts these two alternative actions. As illustrated there, the system will be "restored" to the same state whenever either of the states $(0,s)$ or $(1,-1)$ are again first reached -- these two states

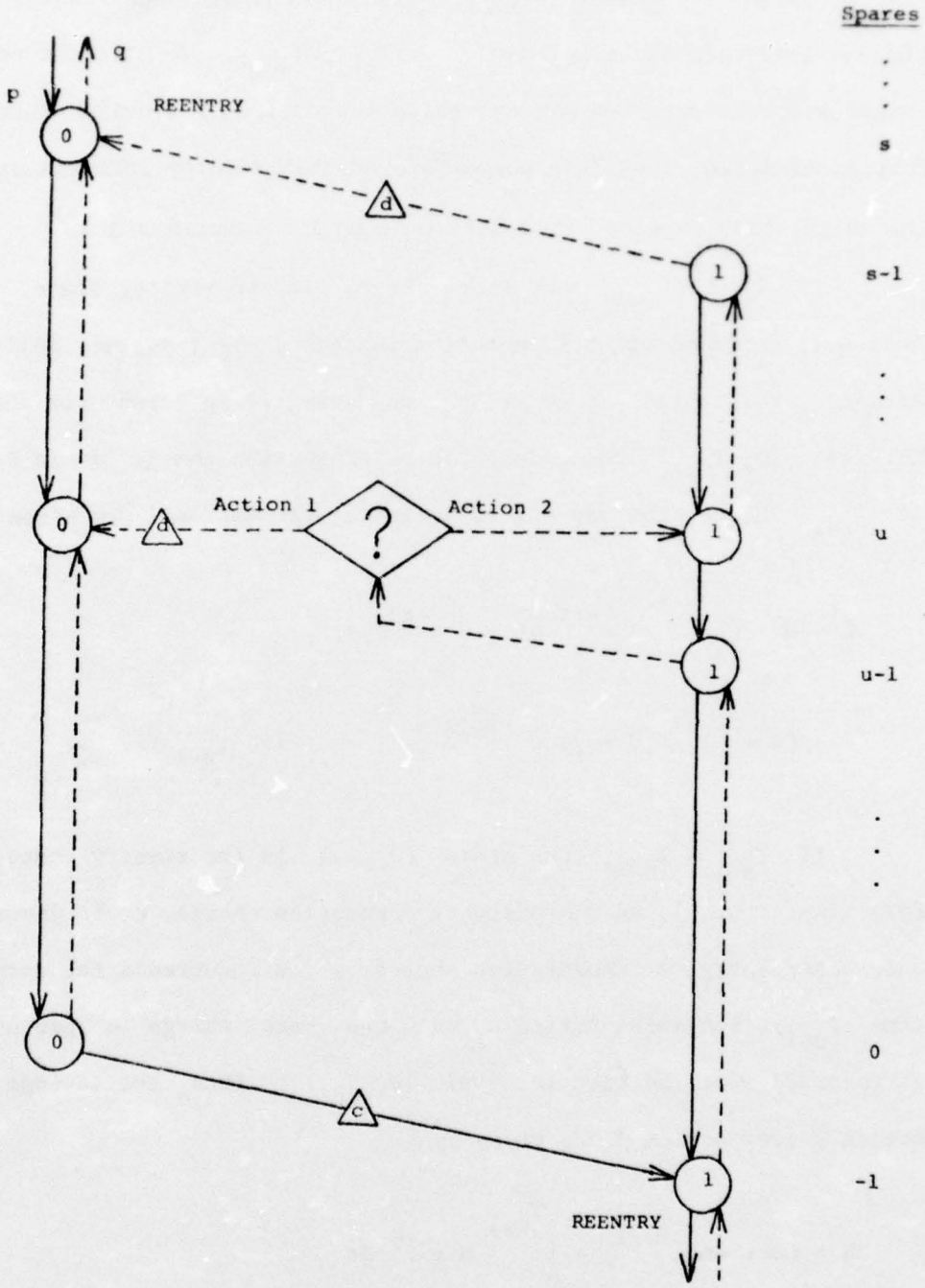


FIGURE 2.5: How long Should a Disconnection be Deferred?

will be called the reentry states. The corresponding incremental reentry time is obviously distributed as $\min\{T_{u+1}, T_{u-s}\}$. To make the economic comparison, the expected savings which accrue from following Action 2 instead of Action 1 will be computed over this reentry interval of time. The calculation is done by conditioning on the reentry state.

If $T_{u-s} < T_{u+1}$, then state $(0,s)$ is the reentry state. Following Action 1, an immediate disconnection charge d is incurred; following Action 2, the rental charge π per unit-time is incurred over the interval $[0, T_{u-s}]$ and a deferred disconnection charge d is incurred at T_{u-s}. Hence, the savings of Action 2 over Action 1 are given by

$$S = d - \left(\int_0^{T_{u-s}} \pi e^{-rt}\, dt + d\, e^{-rT_{u-s}}\right)$$

$$= \left(d - \frac{\pi}{r}\right) - \left(d - \frac{\pi}{r}\right) e^{-rT_{u-s}}, \qquad \text{if } T_{u-s} < T_{u+1}. \qquad (2.1)$$

If $T_{u+1} < T_{u-s}$, then state $(1, -1)$ is the reentry state. Following Action 1, an immediate disconnection charge d is incurred and subsequently, a reconnection charge c is incurred a the reentry time T_{u+1}. Following Action 2, only the rental charge π per unit-time is incurred over the time interval $[0, T_{u+1}]$. Thus, the savings of Action 2 over Action 1 are given by

$$S = \left(d + ce^{-rT_{u+1}}\right) - \int_0^{T_{u+1}} \pi e^{-rt}\, dt$$

$$= \left(d - \frac{\pi}{r}\right) + \left(c + \frac{\pi}{r}\right) e^{-rT_{u+1}}, \qquad \text{if } T_{u+1} < T_{u-s}. \qquad (2.2)$$

Using (2.1) and (2.2), the expected savings of Action 2 over Action 1 are given by

$$
\begin{aligned}
ES &= E[S|T_{u-s} < T_{u+1}]\ P\{T_{u-s} < T_{u+1}\} + E[S|T_{u+1} < T_{u-s}]\ P\{T_{u+1} < T_{u-s}\} \\
&= E[(d - \tfrac{\pi}{r}) - (d - \tfrac{\pi}{r})\ e^{-rT_{u-s}}|T_{u-s} < T_{u+1}]\ P\{T_{u-s} < T_{u+1}\} \\
&\quad + E[(d - \tfrac{\pi}{r}) + (c + \tfrac{\pi}{r})\ e^{-rT_{u+1}}|T_{u+1} < T_{u-s}]\ P\{T_{u+1} < T_{u-s}\} \\
&= (d - \tfrac{\pi}{r}) + (c + \tfrac{\pi}{r})\ E[e^{-rT_{u+1}}|T_{u+1} < T_{u-s}]\ P\{T_{u+1} < T_{u-s}\} \\
&\quad - (d - \tfrac{\pi}{r})\ E[e^{-rT_{u-s}}|T_{u-s} < T_{u+1}]\ P\{T_{u-s} < T_{u+1}\} \\
&= (d - \tfrac{\pi}{r}) + (c + \tfrac{\pi}{r})\ h(u+1,\ u-s) - (d - \tfrac{\pi}{r})\ h(u-s,\ u+1) \\
&= H(u,s)\ . \qquad (2.3)
\end{aligned}
$$

Thus, $H(u,s)$ represents the expected savings accruing from a single disconnection deferral up to spares level s when an opportunity to disconnect into state $(0,u)$ is at hand. In particular, the expected savings accruing from the disconnection deferral previously illustrated in Figure 2.3 is given by $H(0,s)$.

Lemma 2.1.

$$H(u,s) = H(u,v) + H(v,s)\ h(u-v,\ u+1), \quad 0 \le u < v < s\ .$$

Proof. The reentry states are $(0,s)$ and $(1,-1)$, with respective reentry states T_{u-s} and T_{u+1}.

Case 1: $T_{u-v} < T_{u+1}$. In this case, spares level v is reached prior to reentry since $v < s$ implies $T_{u-v} < T_{u-s}$. Since $0 \le u < v < s$,

$$T_{u-s} \overset{D}{=} T_{u-v} + T'_{v-s} \qquad \text{and} \qquad T_{u+1} \overset{D}{=} T_{u-v} + T'_{v+1} ,$$

where $\{T'_i\}$ is a sequence of random variables independent of $\{T_i\}$ and with distributions identical to those of $\{T_i\}$.

Case 1a: $T'_{v-s} < T'_{v+1}$. $(\Rightarrow T_{u-s} < T_{u+1})$. In this subcase, reentry in state $(0,s)$ occurs at reentry time $T_{u-v} + T'_{v-s}$. Proceeding as before, we have

$$S = (d - \tfrac{\pi}{r}) - (d - \tfrac{\pi}{r})\, e^{-rT'_{v-s}}\, e^{-rT_{u-v}} ,$$

$$\text{if } T_{u-v} < T_{u+1} \text{ and } T'_{v-s} < T'_{v+1} . \qquad (2.4)$$

Case 1b: $T'_{v+1} < T'_{v-s}$. $(\Rightarrow T_{u+1} < T_{u-s})$. In this subcase, reentry into state $(1, -1)$ occurs at reentry time $T_{u-v} + T'_{v+1}$. Proceeding as before, we have

$$S = (d - \tfrac{\pi}{r}) + (c + \tfrac{\pi}{r})\, e^{-rT'_{v+1}}\, e^{-rT_{u-v}} ,$$

$$\text{if } T_{u-v} < T_{u+1} \text{ and } T'_{v+1} < T'_{v-s} . \qquad (2.5)$$

Case 2: $T_{u+1} < T_{u-v}$. In this case, reentry into state $(1, -1)$ occurs at reentry time T_{u+1}, without an intermediate visit to spares level v. Hence, as before,

$$S = (d - \frac{\pi}{r}) + (c + \frac{\pi}{r})\, e^{-rT_{u+1}}, \qquad \text{if } T_{u+1} < T_{u-v}. \quad (2.6)$$

For notational convenience, denote the pertinent events by $\mathcal{E}_1 = \{T_{u-v} < T_{u+1}\}$, $\mathcal{E}_{1a} = \{T'_{v-s} < T'_{v+1}\}$, $\mathcal{E}_{1b} = \{T'_{v+1} < T'_{v-s}\}$, and $\mathcal{E}_2 = \{T_{u+1} < T_{u-v}\}$. Note that $P\{\mathcal{E}_{1a} \cup \mathcal{E}_1 | \mathcal{E}_1\} = P\{\mathcal{E}_{1a}\}$, $P\{\mathcal{E}_1 \cup \mathcal{E}_{1b} | \mathcal{E}_1\} = P\{\mathcal{E}_{1b}\}$ and $P\{\mathcal{E}_{1a}\} + P\{\mathcal{E}_{1b}\} = 1$. Also denote $a = d - \pi/r$ and $b = c + \pi/r$. Rewritting (2.4) - (2.6) in new notation gives

$$S = \begin{cases} a - a\, e^{-rT'_{v-s}}\, e^{-rT_{u-v}}, & \text{if } \mathcal{E}_1 \cup \mathcal{E}_{1a} \\ a + b\, e^{-rT'_{v+1}}\, e^{-rT_{u-v}}, & \text{if } \mathcal{E}_1 \cup \mathcal{E}_{1b} \\ a + b\, e^{-rT_{u+1}}, & \text{if } \mathcal{E}_2 . \end{cases}$$

Thus,

$$H(u,s) = ES$$

$$= E[a - a\, e^{-rT'_{v-s}}\, e^{-rT_{u-v}} | \mathcal{E}_1 \cup \mathcal{E}_{1a}]\ P\{\mathcal{E}_1 \cup \mathcal{E}_{1a}\}$$

$$+ E[a + b\, e^{-rT'_{v+1}}\, e^{-rT_{u-v}} | \mathcal{E}_1 \cup \mathcal{E}_{1b}]\ P\{\mathcal{E}_1 \cup \mathcal{E}_{1b}\}$$

$$+ E[a + b\, e^{-rT_{u+1}} | \mathcal{E}_2]\ P\{\mathcal{E}_2\}$$

$$= E\left[\begin{array}{l} E[a - a\, e^{-rT'_{v-s}}\, e^{-rT_{u-v}} | \mathcal{E}_{1a}]\ P\{\mathcal{E}_{1a}\} \\ + \\ E[a + b\, e^{-rT'_{v+1}}\, e^{-rT_{u-v}} | \mathcal{E}_{1b}]\ P\{\mathcal{E}_{1b}\} \end{array} \middle| \mathcal{E}_1 \right]\ P\{\mathcal{E}_1\}$$

$$+ E[a + b\, e^{-rT_{u+1}} | \mathcal{E}_2]\ P\{\mathcal{E}_2\}$$

$$= E\left[\begin{array}{l} \left\{\begin{array}{l} E[a - a\, e^{-rT'_{v-s}} | \mathcal{E}_{1a}]\ P\{\mathcal{E}_{1a}\} \\ + \\ E[a + b\, e^{-rT'_{v+1}} | \mathcal{E}_{1b}]\ P\{\mathcal{E}_{1b}\} \end{array}\right\} e^{-rT_{u-v}} \\ + (a - a\, e^{-rT_{u-v}})\ (P\{\mathcal{E}_{1a}\} + P\{\mathcal{E}_{1b}\}) \end{array} \middle| \mathcal{E}_1 \right]\ P\{\mathcal{E}_1\}$$

$$+ E[a + b\, e^{-rT_{u+1}} | \mathcal{E}_2]\ P\{\mathcal{E}_2\}$$

$$= E[H(v,s)\ e^{-rT_{u-v}} | \mathcal{E}_1]\ P\{\mathcal{E}_1\}$$

$$+ E[a - a\, e^{-rT_{u-v}} | \mathcal{E}_1]\ P\{\mathcal{E}_1\} + E[a + b\, e^{-rT_{u+1}} | \mathcal{E}_2]\ P\{\mathcal{E}_2\}$$

$$= H(v,s)\ h(u-v,\ u+1) + H(u,v)\ . \qquad \square$$

Definition 2.4. Let s^* denote the greatest supremum of $H(0,s)$:

$$s^* = \sup\{s' : H(0,s') = \sup_{s \geq 0} H(0,s)\}.$$

Lemma 2.2. $s^* < +\infty \Longleftrightarrow d < \pi/r$.

Proof. The proof of necessity ($\Rightarrow$) is by contradiction. Suppose $d \geq \pi/r$. Then, for $s \geq 1$,

$$H(s-1,\ s) = (d - \tfrac{\pi}{r}) + (c + \tfrac{\pi}{r})\ h(s,-1) - (d - \tfrac{\pi}{r})\ h(-1,s)$$

$$= (d - \tfrac{\pi}{r})\ (1 - h(-1,\ s)) + (c + \tfrac{\pi}{r})\ h(s,\ -1) > 0\ ,$$

since, by Definition 2.3, $0 < h(u,v) < 1$ for $u < 0 < v$ and $v < 0 < u$. Thus, using Lemma 2.1,

$$H(0,s) = H(0,\ s-1) + H(s-1,\ s)\ h(1-s,\ 1) > H(0,\ s-1), \quad s \geq 1,$$

since $0 < h(1-s,\ 1) < 1$ for $s \geq 1$. Hence, $d \geq \frac{\pi}{r} \Rightarrow s^* \to +\infty$, a contradiction. Thus, the original supposition was false and we must conclude $d < \frac{\pi}{r}$.

To prove sufficiency ($\Leftarrow$), suppose $d < \frac{\pi}{r}$. Then

$$H(s-1,\ s) = (d - \tfrac{\pi}{r})\ (1 - h(-1,s)) + (c + \tfrac{\pi}{r})\ h(s,\ -1)$$

$$= (d - \tfrac{\pi}{r})\ (1 - a_s P_s) + (c + \tfrac{\pi}{r})\ b_s(1-P_s)\ , \qquad (2.7)$$

where $a_s = E[e^{-rT_{-1}} | T_{-1} < T_s]$, $b_s = E[e^{-rT_s} | T_s < T_{-1}]$, and $P_s = P\{T_{-1} < T_s\}$. Since $\lambda, p, q > 0$, it follows that a_s decreases, P_s increases and b_s decreases. Furthermore, $P_s \to 1$ as $s \to +\infty$. Since $0 < b_s < 1$, $b_s(1-P_s) \to 0$ as $s \to +\infty$. As $(c + \frac{\pi}{r})\, b_s(1-P_s) \to 0$, $\exists\ s'$ such that $0 < (c + \frac{\pi}{r})\, b_s(1-P_s) < (\frac{\pi}{r} - d)\,(1-a_1)$, for all $s \geq s'$. Hence, from (2.7), $H(s-1, s) < 0$, for all $s \geq s'$. Thus, using Lemma 2.1, we conclude that $H(0,s) < H(0, s-1)$, for all $s \geq s'$ and it follows that $s^* < s' < +\infty$. □

The economic interpretation of Lemma 2.2 is simple. The lemma states that disconnection will be advantageous if and only if the disconnection charge is less than the cost of renting a module over an infinite horizon. Henceforth, it will always be assumed that $d < \frac{\pi}{r}$.

The next lemma will prove to be quite important from a computational standpoint (see Section 2.8).

Lemma 2.3. s^* is independent of X, K, N, $g(\cdot)$, r_1.

Proof. Since T_{-1} and T_s $(s \geq 0)$ are dependent only on the demand process, $H(0,s)$ is completely determined by the demand process and the parameters c, d, π and r. □

Lemma 2.4. $H(s, s^*) \geq 0$, $s < s^*$ and $H(s^*, s) < 0$, $s > s^*$.

Proof. In view of Lemma 2.1 and Definition 2.3, anything to the contrary would contradict Definition 2.4. □

The next two lemmas establish bounds on the behavior of $h(\cdot,\cdot)$ with regard to each of its arguments.

Lemma 2.5.

$$h(u,s+1) = h(u,s) + E[e^{-rT_u} | T_s < T_u < T_{s+1}]\ P\{T_s < T_u < T_{s+1}\}\ ,$$

$$u < 0,\ s \geq 0.$$

Proof.

$$\begin{aligned} h(u,s+1) &= E[e^{-rT_u} | T_u < T_{s+1}]\ P\{T_u < T_{s+1}\} \\ &= E[e^{-rT_u} | T_u < T_s,\ T_u < T_{s+1}]\ P\{T_u < T_s,\ T_u < T_{s+1}\} \\ &\quad + E[e^{-rT_u} | T_s < T_u < T_{s+1}]\ P\{T_s < T_u < T_{s+1}\} \\ &= E[e^{-rT_u} | T_u < T_s]\ P\{T_u < T_s\} \\ &\quad + E[e^{-rT_u} | T_s < T_u < T_{s+1}]\ P\{T_s < T_u - T_{s+1}\} \\ &= h(u,s) + E[e^{-rT_u} | T_s < T_u < T_{s+1}]\ P\{T_s < T_u < T_{s+1}\}\ . \end{aligned}$$

The next-to-last equality follows since $\{T_u < T_s\} \Rightarrow \{T_u < T_{s+1}\}$. □

Lemma 2.6.

$$h(u,s) > h(u+1,s) + E[e^{-rT_u} | T_u < T_s < T_{u+1}]\ P\{T_u < T_s < T_{u+1}\}\ ,$$

$$u \geq 0,\ s < 0\ .$$

Proof.

$$\begin{aligned}
h(u,s) &= E[e^{-rT_u} | T_u < T_s]\ P\{T_u < T_s\} \\
&= E[e^{-rT_u} | T_u < T_s,\ T_{u+1} < T_s]\ P\{T_u < T_s,\ T_{u+1} < T_s\} \\
&\quad + E[e^{-rT_u} | T_u < T_s < T_{u+1}]\ P\{T_u < T_s < T_{u+1}\} \\
&= E[e^{-rT_u} | T_{u+1} < T_s]\ P\{T_{u+1} < T_s\} \\
&\quad + E[e^{-rT_u} | T_u < T_s < T_{u+1}]\ P\{T_u < T_s < T_{u+1}\} \\
&> E[e^{-rT_{u+1}} | T_{u+1} < T_s]\ P\{T_{u+1} < T_s\} \\
&\quad + E[e^{-rT_u} | T_u < T_s < T_{u+1}]\ P\{T_u < T_s < T_{u+1}\} \\
&= h(u+1,s) + E[e^{-rT_u} | T_u < T_s < T_{u+1}]\ P\{T_u < T_s < T_{u+1}\}\ .
\end{aligned}$$

The third equality follows from $\{T_{u+1} < T_s\} \Rightarrow \{T_u < T_s\}$ for $u \geq 0$. The inequality follows since: (i) for $u \geq 0$, $T_{u+1} > T_u \Rightarrow e^{-rT_u} > e^{-rT_{u+1}}$ (with probability 1); and (ii) $P\{T_{u+1} < T_s\} > 0$ for $u \geq 0$, $s < 0$. □

The economic interpretation of the following theorem is that the marginal savings of disconnection deferrals are decreasing.

Theorem 2.7. $H(s, s+1)$ is decreasing, $s \geq 0$.

Proof.

$$H(s, s+1) = (d - \tfrac{\pi}{r}) + (c + \tfrac{\pi}{r})\, h(s+1, -1) - (d - \tfrac{\pi}{r})\, h(-1, s+1) ,$$

$$H(s-1, s) = (d - \tfrac{\pi}{r}) + (c + \tfrac{\pi}{r})\, h(s, -1) - (d - \tfrac{\pi}{r})\, h(-1, s) .$$

Thus,

$$H(s, s+1) - H(s-1, s) = (c + \tfrac{\pi}{r})\, (h(s+1, -1) - h(s, -1))$$

$$- (d - \tfrac{\pi}{r})\, (h(-1, s+1) - h(-1, s)) .$$

To prove the theorem, it suffices to show that the right-hand side of the above expression is negative. Using Lemmas 2.6 and 2.5 gives

$$h(s,-1) - h(s+1, -1) > E[e^{-rT_s} | T_s < T_{-1} < T_{s+1}]\, P\{T_s < T_{-1} < T_{s+1}\}$$

$$> E[e^{-rT_{-1}} | T_s < T_{-1} < T_{s+1}]\, P\{T_s < T_{-1} < T_{s+1}\}$$

$$= h(-1, s+1) - h(-1, s) .$$

Also, $(c + \tfrac{\pi}{r}) \geq -(d - \tfrac{\pi}{r}) = (\tfrac{\pi}{r} - d) > 0$. Combining this with the above inequality yields

$$(c + \tfrac{\pi}{r})\, (h(s, -1) - h(s+1, -1)) > -(d - \tfrac{\pi}{r})\, (h(-1, s+1) - h(-1, s)) ,$$

which implies

$$(c + \tfrac{\pi}{r})\,(h(s+1, -1) - h(s-1)) - (d - \tfrac{\pi}{r})\,(h(-1, s+1) - h(-1, s)) < 0 .$$

□

Definition 2.5. $H(0,\cdot)$ will be said to be strictly unimodal about $[s', s'']$ if: (i) $H(0,s) < H(0, s+1)$, $s < s'$; (ii) $H(0,s) \leq H(0, s+1)$, $s' \leq s < s''$; and (iii) $H(0,s+1) < H(0,s)$, $s \geq s''$.

The following theorem will be important later with regard to computing s^*.

Theorem 2.8. $H(0,\cdot)$ is strictly unimodal about $[s^*-1, s^*]$.

Proof. From Lemma 2.4, $H(s^*-1, s^*) \geq 0 > H(s^*, s^*+1)$. Therefore, by Theorem 2.7, $H(s-1, s) > 0$ for $s < s^*$ and $H(s, s+1) < 0$ for $s \geq s^*$. Hence, using Lemma 2.1 and the fact that $h(-s,1) > 0$ for all $s \geq 0$,

$$H(0, s+1) = H(0,s) + H(s, s+1)\, h(-s,1) \begin{cases} > H(0,s), & s < s^*-1 \\ \geq H(0, s^*-1), & s = s^*-1 \\ < H(0,s), & s \geq s^* \end{cases}$$ □

Thus, the maximizing set of $H(0,\cdot)$ contains either the single component s^* or the two components s^*-1 and s^*. The optimal of the constant s-policy with parameter s^* will now be established.

Lemma 2.9. If ω is a feasible disconnection policy and if, for any $n \geq 1$, $\inf\{i : \omega(n, -(n-1)L + i) = 1\} > s^*$, then ω is not optimal.

Proof. Suppose ω and n are such that

$$s = \inf\{i : \omega(n, -(n-1)L + i) = 1\} > s^* .$$

To demonstrate the non-optimality of ω, we construct an alternative policy ω' (possibly nonstationary) realizing expected costs strictly less than those of ω. Specifically, define ω' as follows:

(a) $$\omega'(n,k) = \begin{cases} 0, & k < -(n-1)L + s^* \\ 1, & k = -(n-1)L + s^* \end{cases}$$

(b) Whenever ω' disconnects from $(n, -(n-1)L + s^*-1)$ to $(n-1, -(n-1)L + s^*)$, ω' makes no further disconnections until (just prior to) first reaching either (reentry) state $(n, -(n-1)L-1)$ or (reentry) state $(n-1, -(n-1)L+s)$.

(c) ω' is otherwise identical to ω.

If $s = +\infty$, then (reentry) state $(n-1, -(n-1)L + s)$ is interpreted as an unreachable state in (b). If $s < +\infty$ and $\omega(n-1, -(n-1)L+s) = 1$, then (b) and (c) are interpreted to mean that $\omega'(n-1, -(n-1)L+s) = 1$.

Because of (b), ω' need not be stationary. However, we can still represent the two alternative strategies in a transition diagram by defining "dummy" states $\{(n-1,k)', -(n-1)L < k \leq -(n-1)L + s\}$, as shown

in Figure 2.6. Given this enlargement of the state space, ω and ω' can be represented by

$$\omega(n\text{-}1,\ (n\text{-}1)L + s)' = 1\ ,$$

$$\omega'(n\text{-}1,\ k)' = \begin{cases} 0, & -(n\text{-}1)L \leq k < -(n\text{-}1)L + s \\ 1, & k = -(n\text{-}1)L + s \end{cases}$$

and

$$\omega'(i,k) = \begin{cases} \omega(i,k), & i \neq n \\ 0\ , & i = n,\ k < -(n\text{-}1)L + s^* \\ 1\ , & i = n,\ k = -(n\text{-}1)L + s^* \\ \omega(n,k), & i = n,\ k > -(n\text{-}1)L + s^*\ , \end{cases}$$

where it is understood that ω' disconnects from $(n,\ -(n\text{-}1)L + s^* - 1)$ to $(n\text{-}1,\ -(n\text{-}1)L + s^*)'$.

To compare ω and ω', we suppose that the system has been in state $(n,\ -(n\text{-}1)L + s^* - 1)$ and that a demand decrease now occurs. As illustrated in Figure 2.6, policy ω' immediately disconnects, moving to state $(n\text{-}1,\ -(n\text{-}1)L + s^*)'$. Under policy ω, the disconnection is deferred until $s\text{-}s^*$ additional demand decreases occur. At time $T = \min\{T_{s^*+1},\ T_{s^*-s}\}$, the system will be "restored" to identical condition regardless of which strategy is chosen. Hence, any expected savings of ω over ω' must accrue over the time intervals of the form $[0,T]$ (since ω' is identical to ω elsewhere).

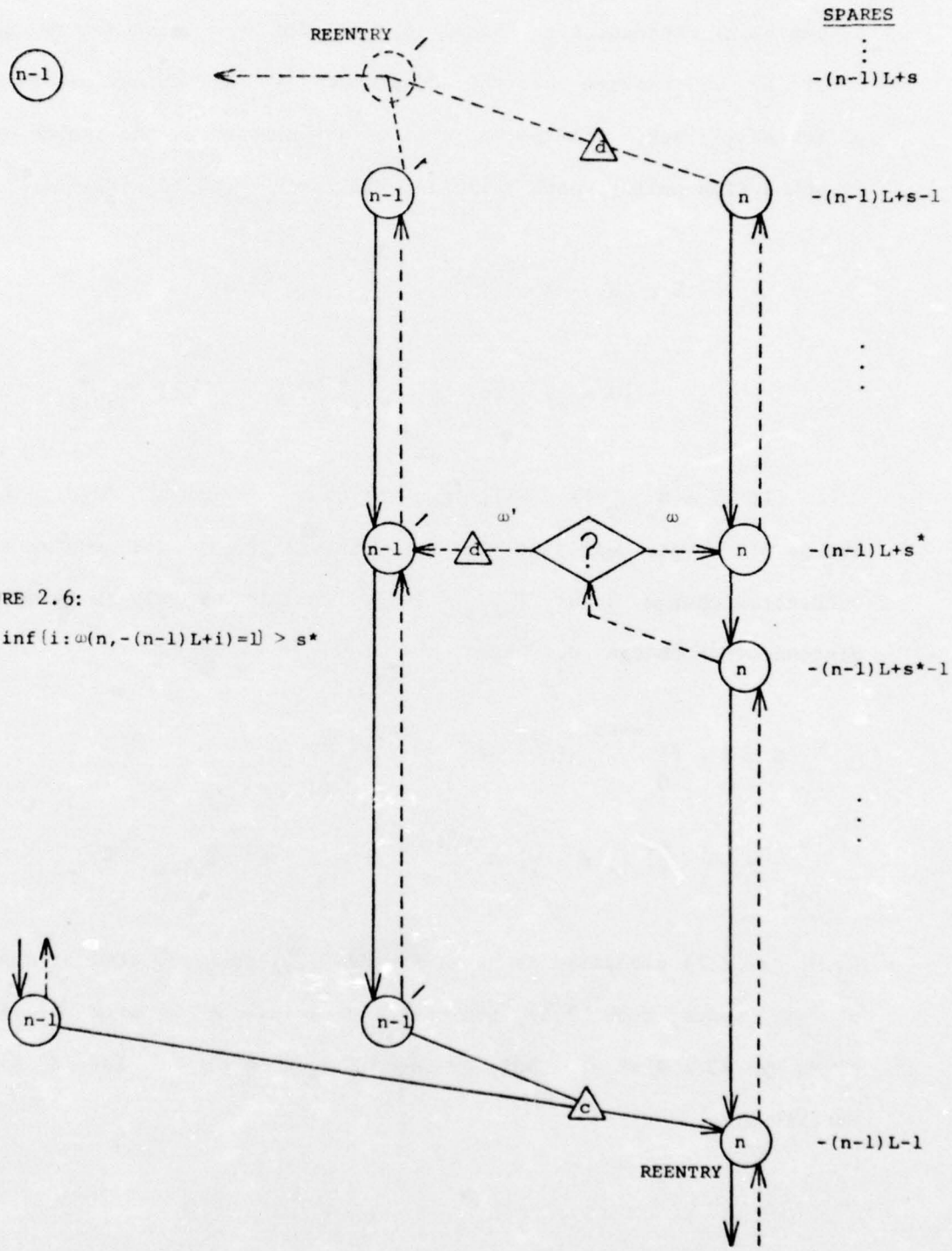


FIGURE 2.6:

$s = \inf\{i: \omega(n,-(n-1)L+i)=1\} > s^*$

If $T = T_{s^*+1}$, then policy ω incurs an additional rental charge of π per unit-time over the interval $[0,T]$, while policy ω' incurs an immediate disconnection charge d, followed by a reconnection charge c at T_{s^*+1}. (Notice that the adjustments $-\gamma_1|k|$ do not enter in the differential, since the spares level is not altered by the choice of disconnection policy; both policies will incur equal adjustments.) Hence,

$$S = (d + ce^{-rT_{s^*+1}}) - \int_0^{T_{s^*+1}} \pi e^{-rt}\, dt$$

$$= (d - \tfrac{\pi}{r}) + (c + \tfrac{\pi}{r})\, e^{-rT_{s^*+1}}, \qquad \text{if } T_{s^*+1} < T_{s^*-s} \quad (2.8)$$

If $T = T_{s^*-s}$ $(s < +\infty)$, then policy ω incurs an additional rental charge of π per unit-time over the interval $[0,T]$ followed by a disconnection charge d at T_{s^*-s}. Policy ω' incurs only an immediate disconnection charge d. Hence,

$$S = d - (\int_0^{T_{s^*-s}} \pi e^{-rt}\, dt + de^{-sT_{s^*-s}})$$

$$= (d - \tfrac{\pi}{r}) - (d - \tfrac{\pi}{r})\, e^{-rT_{s^*-s}}, \qquad \text{if } T_{s^*-s} < T_{s^*+1} \quad (2.9)$$

(2.8) is (2.2) evaluated at $u = s^*$ and (2.9) is (2.1) evaluated at $u = s^*$. Hence, from (2.3), the expected savings of ω over ω' are given by $ES = H(s^*,s)$. But, by Lemma 2.4, $H(s^*,s) < 0$ for $s^* < s < +\infty$. Furthermore,

$$H(s^*,\infty) = (d - \tfrac{\pi}{r}) + (c + \tfrac{\pi}{r})\, E[e^{-rT_{s^*+1}}] \leq H(s^*,s) \qquad \text{for all} \quad s > s^*,$$

since (2.9) is positive for all $s > s^*$; hence, $H(s^*, \infty) < 0$. Thus, ω realizes (strictly) negative expected savings over ω'; that is, ω' realizes expected costs strictly less than those of ω. Hence, ω is not optimal. □

Lemma 2.10. There exists an optimal disconnection policy ω^* with

$$\omega^*(1,k) = \begin{cases} 0, & k < s^* \\ 1, & k = s^* \end{cases}.$$

Proof. Referring to Figure 2.2, suppose that the system has been in state $(1, -1)$ and that a departure now occurs (i.e., $k \to 0$ from -1). Compare the two alternative strategies of either an immediate disconnection or a disconnection deferral until $s \geq 0$ units of spare capacity will remain. The savings accruing from the deferral strategy is $H(0,s)$, $s \geq 0$. We can do no better than maximize this savings with $H(0,s^*)$, by assigning $\omega^*(1,\cdot)$ as above. □

Theorem 2.11. There exists an optimal disconnection policy ω^* with

$$\omega^*(n,k) = \begin{cases} 0, & k < -(n-1)L + s^* \\ 1, & k = -(n-1)L + s^* \end{cases}; \quad n = 1, 2, \ldots . \qquad (2.10)$$

Proof (induction). By Lemma 2.10, the theorem is true for $n = 1$ and we now assume that, in general, it is true over $1, 2, \ldots, n-1$, for some $n-1 \geq 1$.

Let ω be an arbitrary policy agreeing with (2.10) over $1, 2, \ldots, n-1$. Let

$$s = \inf\{i : \omega(n, -(n-1)L + i) = 1\} .$$

By Lemma 2.9, it suffices to assume that $s \leq s^*$. Let ω' be a policy agreeing with ω everywhere except for $\omega'(n, -(n-1)L + i) = 0$, $0 \leq i < s^*$ and $\omega'(n, -(n-1)L + s^*) = 1$. We shall show that the expected savings of ω' over ω are nonnegative. If $s = s^*$, then $\omega' \equiv \omega$ and there is nothing to prove; hence assume $s < s^*$.

The two alternatives are illustrated in Figure 2.7. Suppose, the system is in state $(n, -(n-1)L + s - 1)$ and that a demand decrease now occurs. Let $T = \min\{T_{s+1}, T_{s-s^*}\}$. Under policy ω, an immediate disconnection to state $(n-1, -(n-1)L + s)$ occurs. Furthermore, under policy ω, the system actual enters state $(n-1, -(n-1)L + s)$ and no further disconnections are made over the time interval $[0,T]$ since, by the induction hypothesis, $\omega(n-1, (n-1)L + i) = \omega(n-1, (n-2)L + (i-L)) = 0$ for $i < L+s^*$ (and $s^* < L+s^*$). Under policy ω', the disconnection is deferred until s^* units of spare capacity will remain. At time T, the system is restored at one of the two reentry states $(n, -(n-1)L - 1)$ or $(n-1, (n-1)L + s^*)$ (regardless of which policy is chosen) depending upon whether $T = T_{s+1}$ or $T = T_{s-s^*}$, respectively.

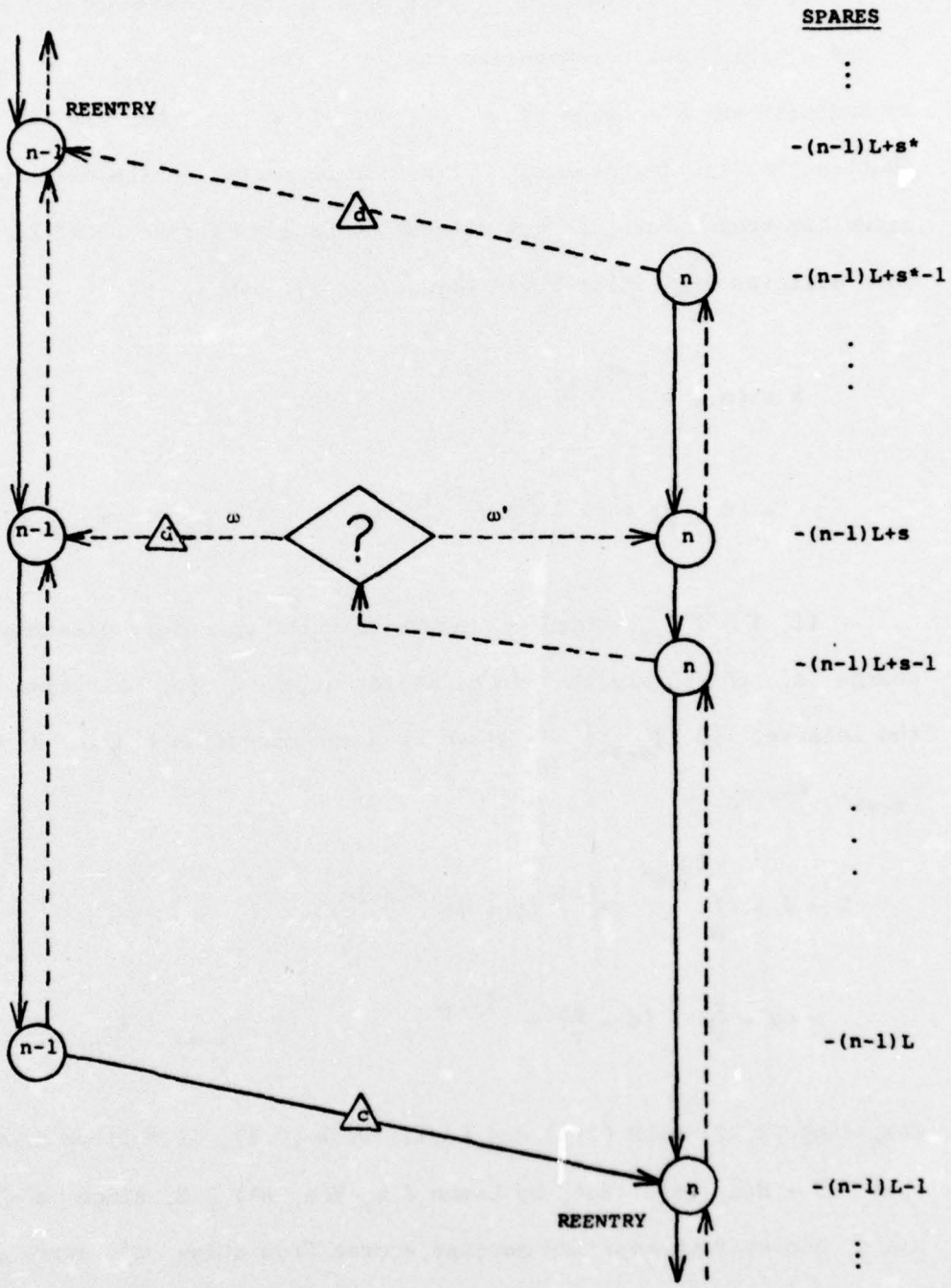


Figure 2.7: $s = \min\{i:\omega(n, -(n-1)L+i) = 1\} < s^*$

If $T = T_{s+1}$, then ω incurs an immediate disconnection charge d and a subsequent reconnection charge c at time T_{s+1}. ω' incurs an addition rental charge of π per unit-time over the interval $[0,T]$. (Notice that the adjustments $-\gamma_1|k|$ do not enter in the differential, since the spares level is not altered by choice of disconnection policy; both policies will incur equal adjustments.) Hence,

$$S = (d + ce^{-rT_{s+1}}) - \int_0^{T_{s+1}} \pi e^{-rt}\, dt$$

$$= (d - \tfrac{\pi}{r}) + (c + \tfrac{\pi}{r})\, e^{-rT_{s+1}}, \qquad \text{if } T_{s+1} < T_{s-s^*}. \quad (2.11)$$

If $T = T_{s-s^*}$, then ω incurs only the immediate disconnection charge d. ω' incurs the rental differential π per unit-time over the interval $[0, T_{s-s^*}]$ followed by a disconnection charge d at T_{s-s^*}. Hence,

$$S = d - (\int_0^{T_{s-s^*}} \pi e^{-rt}\, dt + de^{-rT_{s-s^*}})$$

$$= (d - \tfrac{\pi}{r}) - (d - \tfrac{\pi}{r})\, e^{-rT_{s-s^*}}, \qquad \text{if } T_{s-s^*} < T_{s+1}. \quad (2.12)$$

Comparing (2.12) with (2.1) and (2.11) with (2.2), it follows from (2.3) that $ES = H(s, s^*)$. But, by Lemma 2.4, $H(s, s^*) \geq 0$, since $s < s^*$. Hence, nonnegative expected savings accrue from using ω' over ω. Since ω was arbitrary, it follows that the theorem holds over $1, 2, \ldots, n$. The theorem then follows by induction on n. □

Thus, the optimality of constant s-policies is established.

2.4. The Transition Equations for Constant s-Policies, $s \geq 1$

Recall that K denotes an upper limit on the temporary facility capacity that can be used. Since the temporary facilities are modular of size L, only those values of K being nonnegative integer multiples of L are of interest here. Hence, K will be denoted as $K = NL$, where N represents the limiting number of temporary modules allowed.

$C_k^n(X, N, s)$ will denote the expected total discounted cost when the initial spares level is k, the initial number of temporary modules connected is n, the limit on temporary module usage is N, the permanent capacity expansion size is X+1, and when the module disconnection pattern adhers to the constant s-policy having parameter s; $X \geq NL$, $N \geq 0$, $s \geq 1$, $k \geq -NL$, $0 \leq n \leq N$. The constant s-policy having parameter $s = 0$ is not considered here, since that case has previously been treated by Model I of [7].

Similarly, F_k^n will denote the expected incremental discounted costs (until the next arrival or departure) while the spares level has value k, when there are n modules connected. Notice that, by Assumption 2.2, these costs are independent of N. Also, by Assumption 2.1', $F_k^0 = 0$ for $k > 0$ and F_k^n includes an (expected) adjustment term

$$\int_0^\infty \gamma_1 k r^{-1}(1 - e^{-rt})\lambda e^{-\lambda t}\, dt = (\lambda+r)^{-1}\gamma_1 k , \qquad k \leq 0 .$$

Recall that p is the probability of arrival and $q = 1-p$ is the probability of departure. Denote:

$$\alpha = \lambda q(\lambda+r)^{-1} \qquad \text{and} \qquad \beta = \lambda p(\lambda+r)^{-1} \ . \tag{2.13}$$

Also, denote:

$$\Pi = (\lambda+r)^{-1}\,\pi \qquad \text{and} \qquad \Gamma = (\lambda+r)^{-1}\,\gamma_1 \ . \tag{2.14}$$

When a constant s-policy is in use, there are four different types of states: (a) potential module connection states, (b) potential module disconnection states, (c) states where neither connection nor disconnection possibilities exist, and (d) a potential permanent expansion state. The incremental expected costs for each of these state types are derived below.

(a) Potential connection states are of the form (n,k) where $k = -nL$, $0 \le n \le N-1$. At these states, an arrival triggers the connection of an additional module at cost c. Conditioning on the time and the type of the next event gives

$$F_k^n = \int_0^\infty (pc\, e^{-rt} + (n\pi + k\gamma_1)\, r^{-1}(1 - e^{-rt}))\lambda\, e^{-\lambda t}\, dt$$

$$= \beta c + n\Pi + k\Gamma \ , \qquad k = -nL,\ 0 \le n \le N-1 \ .$$

(b) Potential disconnection states are of the form (n,k) where $k = -(n-1)L + s - 1$, $1 \le n \le N$. At these states, a departure triggers the disconnection of a module at cost d. Notice that these states are distinct from those of (a), since $s \ge 1$. Conditioning on the time and the type of the next event gives

$$F_k^n = \int_0^\infty (qd\, e^{-rt} + (n\pi + \min(0,k)\, \gamma_1)\, r^{-k}(1 - e^{-rt}))\lambda\, e^{-\lambda t}\, dt$$

$$= \alpha d + n\Pi + \min(0,k)\Gamma \quad , \qquad k = -(n-1)L + s - 1,\ 1 \le n \le N \ .$$

(c) Intermediate states where there is no potential for either connection or disconnection are of the form (n,k), where $-nL+1 \le k \le -(n-1)L+s-2$, $1 \le n \le N$. These states are distinct from those of (a) and (b). However, in the special case of $s = L = 1$, these states will be nonexistent. Conditioning on the time of the next event gives

$$F_k^n = \int_0^\infty (n\pi + \min(0,k)\, \gamma_1)\, r^{-1}(1 - e^{-rt})\lambda\, e^{-\lambda t}\, dt$$

$$= n\Pi + \min(0,k)\Gamma \quad , \qquad -nL+1 \le k \le -(n-1)L+s-2,\ 1 \le n \le N \ .$$

(d) The potential expansion state is $(N, -NL)$ and an arrival in this case triggers an expansion. Conditioning on the time of the next event gives

$$\int_0^\infty (N\pi - NL\, \gamma_1)\, r^{-1}(1 - e^{-rt})\lambda\, e^{-\lambda t}\, dt$$

$$= N\Pi - NL\Gamma \ .$$

When expansion occurs, there may be extraordinary costs (seperate from the expansion cost) for disposing of the temporary facilities. These costs will be denoted $\phi_e(N)$. For computational reasons, it will prove convenient to write

$$F^N_{-NL} = \beta c + N\Pi - NL\Gamma \quad .$$

It will be demonstrated shortly that this causes no inaccuracies provided that a term "-c" is included in the disposal cost function $\phi_e(\cdot)$. For Example, $\phi_e(N) = Nd-c$ would account for the disconnection of the N modules when permanent expansion occurs, as well as rectifying the non-existent connection charge included in the above expression of F^N_{-NL}. It will also be assumed that $\phi_e(\cdot)$ is nondecreasing and that $c + \phi_e(\cdot)$ is nonnegative (but otherwise arbitrarily given).

Summarizing, the expected incremental costs are given by:

$$F^n_k = \begin{cases} 0 , & k > 0,\ n = 0 \\ \beta c + n\Pi + k\Gamma , & k = -nL,\ n \geq 0 \\ \alpha d + n\Pi + \min(0,k)\Gamma , & k = -(n-1)L+s-1,\ n \geq 1 \\ n\Pi + \min(0,k)\Gamma , & -nL+1 \leq k \leq -(n-1)L+s-2,\ n \geq 1 \end{cases}$$

By conditioning on the time and type of the next event, the cost transition equations have the general form

$$C^n_k(X,N,s) = F^n_k + \int_0^\infty \lambda e^{-\lambda t}\left(qC^{n'}_{k+1}(X,N,s) + pC^{n''}_{k-1}(X,N,s)\right) e^{-rt}\, dt$$

$$= F^n_k + \alpha C^{n'}_{k+1}(X,N,s) + \beta C^{n''}_{k-1}(X,N,s) ,$$

where $n' \in \{n-1,n\}$ and $n'' \in \{n,n+1\}$ depending on whether or not either a connection or a disconnection state is involved. Given the expected incremental costs derived in (a) - (c), the cost transition equations for a constant s-policy are:

$$C_k^o(X,N,s) = \alpha C_{k+1}^o(X,N,s) + \beta C_{k-1}^o(X,N,s), \qquad k > 0 \tag{2.15}$$

$$C_k^n(X,N,s) = F_k^n + \alpha C_{k+1}^n(X,N,s) + \beta C_{k-1}^{n+1}(X,N,s),$$
$$k = -nL,\ 0 \le n \le N-1 \tag{2.16}$$

$$C_k^n(X,N,s) = F_k^n + \alpha C_{k+1}^{n-1}(X,N,s) + \beta C_{k-1}^n(X,N,s),$$
$$k = -(n-1)L+s-1,\ 1 \le n \le N \tag{2.17}$$

$$C_k^n(X,N,s) = F_k^n + \alpha C_{k+1}^n(X,N,s) + \beta C_{k-1}^n(X,N,s),$$
$$-nL+1 \le k \le -(n-1)L+s-1,\ 1 \le n \le N\ . \tag{2.18}$$

Notice that no upper bound is imposed upon k in (2.15). This corresponds to the "untruncated demand assumption" discussed in Chapter 1. Therefore, the comments in [7] (Section 1.7) regarding the viability of this assumption are applicable here.

When $k = -NL$ and an arrival next occurs, a permanent expansion of size $X+1$ $(X \ge NL)$ is undertaken at cost $g(X)$. Whenever this happens, the spares level increases to the value $X-NL \ge 0$ and all

temporary modules are disposed of at cost $\phi_e(N)$. Therefore, conditioning on the next event time and type, the boundary transition equation becomes

$$C^N_{-NL}(X,N,s)$$

$$= F^N_{-NL} + \int_0^\infty \lambda e^{-\lambda t}\{qC^N_{-NL+1}(X,N,s) + p(\phi_e(N) + g(X) + C^o_{X-NL}(X,N,s))\}e^{-rt}\,dt$$

$$= F^N_{-NL} + \alpha C^N_{-NL+1}(X,N,s) + \beta\{\phi_e(N) + g(X) + C^o_{X-NL}(X,N,s)\}\ . \qquad (2.19)$$

Notice that the inclusion of the "-c" term in $\phi_e(N)$ will indeed cancel the nonexistent "βc" charge term of F^N_{-NL}.

The solutions to equations (2.15) - (2.19) are treated in the next four sections. A final note: when $N = 0$, the system of equations consists of only (2.15) and (2.19).

2.5. Solutions for the Transition Equations, $k > 0$, $n = 0$

The special nature of the linear equations (2.15) are well-known [1], [3]. As the solutions to these equations are derived in [7], the corresponding results are simply listed here.

Lemma 2.12. For α and β given by (2.13):

(i) $\alpha > 0$, $\beta > 0$,

(ii) $\alpha+\beta < 1$,

(iii) $\alpha\beta < 1/4$.

Proof. See Lemma 2.1 of [7]. □

Theorem 2.13.

$$C_k^o(X,N,s) = Z^k \, C_0^o(X,N,s) \ , \qquad k \geq 0 \ , \tag{2.20}$$

where

$$Z = \frac{1 - \sqrt{1 - 4\alpha\beta}}{2\alpha} \ . \tag{2.20a}$$

Furthermore, Z is real and $\beta < Z < \alpha+\beta$. Also, $\alpha Z^2 - Z + \beta = 0$.

Proof. See Theorem 2.2 of [7]. □

Corollary 2.14. For $X \geq NL$, $X' \geq N'L$, $s \geq 1$, $s' \geq 1$ and $k \geq 0$,

$$C_0^o(X,N,s) < C_0^o(X',N',s') \iff C_k^o(X,N,s) < C_k^o(X',N',s') \quad .$$

Proof. The result is a direct consequence of the theorem since $Z > \beta > 0$. □

As shown in [7], Z^i is the Laplace transform of T_i, evaluated at the discount rate r, for $i \geq 1$.

Lemma 2.15. $Z^i = E[e^{-rT_i}]$, $i \geq 1$.

Proof. See Lemmas 2.4 and 2.5 of [7]. □

2.6. The Form of the Functionals $\{C_0^o(\cdot,N,s)\}$

When temporary modules are in use, the expected costs behave according to (2.16) - (2.19). Since $X \geq NL$, it follows from Theorem 2.13 that (2.19) may be rewritten as

$$C_{-NL}^N(X,N,s) = F_{-NL}^N + \alpha C_{-NL+1}^N(X,N,s) + \beta(\phi_e(N) + g(X) + Z^{X-NL} C_0^o(X,N,s)) \quad . \tag{2.21}$$

Also, $C_1^o(X,N,s) = ZC_0^o(X,N,s)$, so the first equation of (2.16) can be rewritten as

$$C_0^o(X,N,s) = (1-\alpha Z)^{-1} F_0^o + \beta(1-\alpha Z)^{-1} C_{-1}^1(X,N,s) \quad .$$

Let $X = \hat{X}$, $N = \hat{N}$ and $s = \hat{s} \geq 1$ be fixed. Denote $\hat{C}_k^n = C_k^n(\hat{X},\hat{N},\hat{s})$. Then (2.16) - (2.18) and (2.21) yield a square $(\hat{N}(\hat{s}+L) + 1)$ nonsingular linear system that can be solved to yield the costs $\{\hat{C}_k^n\}$ for the fixed values $\hat{X}$, $\hat{N}$ and $\hat{s}$. By Corollary 2.14, it suffices to compute only $\hat{C}_0^o$.

Obviously, it is impractical to solve such systems for all possible values $(\hat{X}, \hat{N}, \hat{s})$, so some sort of parameterization is necessary. Suppose that the boundary equation (2.21) for the fixed values $(\hat{X}, \hat{N}, \hat{s})$ is replaced by an equation

$$\hat{C}_{-NL}^N = F_{-NL}^N + \alpha\hat{C}_{-NL+1}^N + \beta\hat{C}_d \ ,$$

where $\hat{C}_d$ is a dummy variable. Then the resulting linear system is independent of the value $\hat{X}$ and solving for $\hat{C}_0$ in terms of $\hat{C}_d$ will yield $\hat{C}_0 = \hat{a} + \hat{b}\hat{C}_d$. Making the substitution $\hat{C}_d = g(\hat{X}) + Z^{\hat{X}-\hat{N}L}\hat{C}_0^o + \phi_e(\hat{N})$ then yields $\hat{C}_0^o = (\hat{a} + \hat{b}g(\hat{X}) + \hat{b}\phi_e(N))\ (1 - \hat{b}Z^{\hat{X}-\hat{N}L})^{-1}$. Since this is true for any value $\hat{X}$ chosen, it follows that $C_0^o(X,\hat{N},\hat{s}) = (\hat{a} + \hat{b}g(X) + \hat{b}\phi_e(\hat{N}))$ $(1 - \hat{b}Z^{X-\hat{N}L})^{-1}$, for all $X \geq \hat{N}L$. Hence, using the dummy substitution, the system need only be solved once for given values $(\hat{N},\hat{s})$ to obtain the functional $C_0^o(\cdot,\hat{N},\hat{s})$, which can then be minimized over $[\hat{N}L, \infty)$.

With regard to the parameter N, it should be obvious that the respective linear systems for $\hat{N}$ and $\hat{N}+1$ would be very similar. In fact, the first $\hat{N}(\hat{s}+L)$ equations of the system for $\hat{N}$ are part of the system for $\hat{N}+1$. Hence, it seems reasonable to expect that solving

for the functional $C_0^o(\cdot,\hat{N},\hat{s})$ should aid in solving for the functional $C_0^o(\cdot, \hat{N}+1, \hat{s})$. The next two sections develop procedures for recursively obtaining the functionals $C_0^o(\cdot,N,\hat{s})$ for $N = 0, 1, 2, \ldots$.

With regard to the parameter s, recall that the optimal parameter value s^* is independent of X, N and $g(\cdot)$ (Lemma 2.3). Given the recursions of the next section, this fact is used to develop a simple procedure for determining s^* at the onset of the computations. Once s^* is determined, only the functionals $\{C_0^o(\cdot,N,s^*)\}$ need be computed.

The remainder of this section uses probabilistic arguments to investigate the functional form of $C_0^o(\cdot,N,s)$ that can be anticipated. Let $G_k^o(N,s)$ denote the expected discounted costs until the first expansion, starting from spares level $k \geq 0$ (with no modules initially connected), when N is the limit on temporary module usage and the constant s-policy with parameter s is in use.

Lemma 2.16.

$$G_k^o(N,s) = Z^k \, G_0^o(N,s) \quad , \qquad k \geq 0 \quad . \tag{2.22}$$

Proof. Starting from spares level $k \geq 0$, no costs are incurred until the spares level first becomes nonpositive; the time of this event is given by T_k and $E[e^{-rT_k}] = Z^k$ by Lemma 2.15. The expected costs beyond T_k, until the first expansion, are given by $G_0^o(N,s)$ (discounted relative to T_k). Hence, $G_k^o(N,s) = E[G_0^o(N,s)\, e^{-rT_k}] = G_0^o(N,s)\, E[e^{-rT_k}]$ $= G_0^o(N,s)Z^k$. □

When an expansion occurs, all temporary modules are disconnected and the spares level changes to $X\text{-}NL \geq 0$. Hence, the expected inter-expansion costs are given by

$$G^{o}_{X-NL}(N,s) = Z^{X-NL}\, G^{o}_{0}(N,s) \quad .$$

Theorem 2.17.

$$C^{o}_{0}(X,N,s) = \frac{\phi(N,s) + g(X)}{Z^{-1} - Z^{X}}\, Z^{NL} , \qquad X \geq NL \tag{2.23}$$

where

$$\phi(N,s) = G^{o}_{0}(N,s)\, Z^{-(NL+1)} \geq 0 . \tag{2.23a}$$

Proof. Starting from spares level zero, the time until the first expansion is given by T_{NL+1} and $E[e^{-rT_{NL+1}}] = Z^{NL+1}$, by Lemma 2.15. At time T_{NL+1}, an expansion of size $X+1$ occurs at cost $g(X)$ and the expected costs thereafter are $C^{o}_{X-NL}(X,N,s) = Z^{X-NL}\, C^{o}_{0}(X,N,s)$ (discounted relative to T_{NL+1}). Hence,

$$\begin{aligned} C^{o}_{0}(X,N,s) &= G^{o}_{0}(N,s) + E[(g(X) + Z^{X-NL}\, C^{o}_{0}(X,N,s))\, e^{-rT_{NL+1}}] \\ &= G^{o}_{0}(N,s) + (g(X) + Z^{X-NL}\, C^{o}_{0}(X,N,s))\, E[e^{-rT_{NL+1}}] \\ &= G^{o}_{0}(N,s) + (g(X) + Z^{X-NL}\, C^{o}_{0}(X,N,s))\, Z^{NL+1} \\ &= G^{o}_{0}(N,s) + g(X)\, Z^{NL+1} + Z^{X+1}\, C^{o}_{0}(X,N,s) \end{aligned}$$

$$C_0^o(X,N,s) = \frac{G_0^o(N,s) + g(X)\ Z^{NL+1}}{1 - Z^{X+1}}$$

$$= \frac{G_0^o(N,s)\ Z^{-(NL+1)} + g(X)}{Z^{-1} - Z^X}\ Z^{NL}$$

$$= \frac{\phi(N,s) + g(X)}{Z^{-1} - Z^X}\ Z^{NL} ,$$

where $\phi(N,s)$ is given by (2.23a). By Assumption 2.2, $\Pi > \gamma_1 L$, $c > 0$ and $d > 0$. Hence, all costs incurred are nonnegative. Thus, $G_0^o(N,s) \geq 0$ and it then follows that $\phi(N,s) \geq 0$, since $Z > 0$. □

Rearranging (2.23a) gives $G_0^o(N,s) = \phi(N,s)\ Z^{NL+1} = \phi(N,s)E[e^{-rT_{NL+1}}]$. Similarly, using (2.22) gives

$$G_k^o(N,s) = \phi(N,s)\ Z^{NL+k+1} = E[e^{-rT_{NL+k+1}}] , \quad k \geq 0 .$$

When the initial spares level is $k \geq 0$, the time of the first expansion is T_{NL+k+1}. Thus, in an expectation sense, $\phi(N,s)$ is the equivalent lump sum payment that would be incurred at the first permanent expansion time in lieu of the incremented charges over the time span previous to the expansion. Since the expected inter-expansion costs are $G_{X-NL}^o(N,s)$ $= \phi(N,s)\ E[e^{-T_{X+1}}]$ and the inter-expansion time intervals are (in distribution) of size T_{X+1}, a similar lump sum interpretation of $\phi(N,s)$ can be given with regard to the inter-expansion costs.

Lemma 2.18. $\phi(\cdot,s)$ is nondecreasing, $s \geq 1$.

Proof. The expected costs over the time interval $[0, T_{NL+1})$ are $G_0^o(N,s) - \phi_e(N)\, Z^{NL+1}$, regardless of whether the module limit is N or N+1. The expected costs over the greater (with probability 1) time interval $[0, T_{(N+1)L+1})$ are $G_0^o(N+1,s) - \phi_e(N+1)\, Z^{(N+1)L+1}$. Hence, using (2.23a),

$$G_0^o(N+1,s) - \phi_e(N+1)\, Z^{(N+1)L+1} \geq G_0^o(N,s) - \phi_e(N)\, Z^{NL+1}$$

$$\Rightarrow$$

$$(\phi(N+1,s) - \phi_e(N+1))Z^L \geq \phi(N,s) - \phi_e(N)$$

$$\Rightarrow$$

$$\phi(N+1,s) - \phi_e(N+1) \geq \phi(N,s) - \phi_e(N)$$

$$\Rightarrow$$

$$\phi(N+1,s) \geq \phi(N,s) + \phi_e(N+1) - \phi_e(N) \geq \phi(N,s)\ ,$$

since the disposal costs $\phi_e(\cdot)$ are assumed nondecreasing. □

By Theorem 2.17, the coefficient $\phi(N,s)$ completely determines the functional $C_0^o(\cdot,N,s)$. Hence, the next section focuses on developing a procedure for recursively computing $\phi(N,s)$ over $N = 0, 1, 2, \ldots$.

2.7. Recursive Computation of the Functionals $\{C_0^o(\cdot,N,s),\ N \geq 0\}$

As demonstrated in the previous section, the coefficient $\phi(N,s)$ completely determines the functional $C_0^o(\cdot,N,s)$. For a given value $s \geq 1$, this section provides a procedure for recursively computing the coefficients $\{\phi(N,s),\ N = 0, 1, \ldots\}$ and hence, the functionals $\{C_0^o(\cdot,N,s),\ N = 0, 1, \ldots\}$. The next section provides a procedure for determining the optimal disconnect parameter s^* at the onset of computations. Thus, the algorithms presented here need only be implemented for the given value $s = s^*$.

Definition 2.5. For $m \geq 0$, let $A(m)$ denote the square tridiagonal matrix of dimension $m+1$ with nonzero elements given by

$$A_{ii}(m) = 1, \qquad i = 0, 1, \ldots, m$$

$$A_{i,i+1}(m) = -\beta, \qquad i = 0, 1, \ldots, m-1$$

$$A_{i,i-1}(m) = -\alpha, \qquad i = 1, 2, \ldots, m .$$

Definition 2.6. Let $\bar{C}, \bar{R} \in \mathbb{R}^{m+1}$ and $Q \in \mathbb{R}^{(m+2)\times(m+2)}$. The square linear system

$$Q\begin{pmatrix} C_0 \\ \bar{C} \end{pmatrix} = \begin{pmatrix} R_0 \\ \bar{R} \end{pmatrix}$$

is a Q-system if Q is nonsingular and

$$Q = \begin{bmatrix} 1 & U \\ V & A \end{bmatrix} ,$$

where $A = A(m)$, $U^T \in \mathbb{R}^{m+1}$, $V \in \mathbb{R}^{m+1}$,

$$U_j = \begin{cases} 0 , & j \neq \ell \\ b , & j = \ell \end{cases}$$

and

$$V_i = \begin{cases} a , & i = 0 \\ 0 , & i \neq 0 \end{cases}$$

for some a, b and some $\ell \geq 0$.

The following lemma demonstrates that the solution to a Q-system is determined by A^{-1}.

Lemma 2.19. If the square linear system

$$Q\begin{pmatrix} C_0 \\ \bar{C} \end{pmatrix} = \begin{pmatrix} R_0 \\ \bar{R} \end{pmatrix}$$

is a Q-system, then

$$C_0 = (1 - abA_{\ell 0}^{-1})^{-1} (R_0 - bA_{\ell .}^{-1} \bar{R}) \tag{2.24}$$

and

$$\bar{C} = A^{-1}\bar{R} - aC_0 A_{.0}^{-1} \quad . \tag{2.25}$$

Proof. Partition Q^{-1} as

$$Q^{-1} = \begin{bmatrix} \epsilon & u \\ v & \mathcal{Q} \end{bmatrix} ,$$

where $\epsilon \in \mathbb{R}$, $u^T \in \mathbb{R}^{m+1}$, $v \in \mathbb{R}^{m+1}$ and $\mathcal{Q} \in \mathbb{R}^{(m+1)\times(m+1)}$. Then, from Definition 2.6, $Q^{-1}Q = I$ implies that

$$\epsilon + au_0 = 1 \quad , \tag{2.26}$$

$$\epsilon U + uA = (0, \ldots, 0) \quad . \tag{2.27}$$

(2.27) gives, using Definition 2.6,

$$u = -\epsilon U A^{-1} = -\epsilon b A_{\ell .}^{-1} \quad . \tag{2.28}$$

Thus, $u_0 = -\epsilon b A_{\ell 0}^{-1}$. Substituting this result into (2.26) gives

$$\epsilon = (1 - abA_{\ell 0}^{-1})^{-1} \quad .$$

Hence, from (2.28),

$$u = -(1 - abA_{\ell 0}^{-1})^{-1}\, bA_{\ell .}^{-1} \quad ,$$

and thus,

$$C_0 = Q_{0.}^{-1}\begin{pmatrix} R_0 \\ \bar{R} \end{pmatrix} = (\epsilon,\ u)\begin{pmatrix} R_0 \\ \bar{R} \end{pmatrix}$$

$$= (1 - abA_{\ell 0}^{-1})\ (R_0 - bA_{\ell .}^{-1}\ \bar{R}) \quad .$$

Using Definition 2.6,

$$aC_0\ e(0) + A\bar{C} = \bar{R} \ ,$$

where $e(0) = (1, 0, \ldots, 0)^T \in \mathbb{R}^{m+1}$. Hence,

$$\bar{C} = A^{-1}(\bar{R} - aC_0\ e(0))$$

$$= A^{-1}\bar{R} - aC_0\ A_{.0}^{-1} \quad . \qquad \square$$

The next lemma treats the special case $N = 0$.

Lemma 2.20.

$$\phi_a(0,s) = \phi_a(0,s) + \phi_e(0) \ ,$$

where

$$\phi_a(0,s) = \beta^{-1}\ F_0^o \ , \qquad s \geq 0 \ .$$

Proof. When $N = 0$, the transition equations are independent of s (since temporary facilities are never used):

$$C_k^o = Z^k C_0^o , \qquad k \geq 0$$

$$-\alpha C_1^o + C_0^o = F_0^o + \beta\{g(X) + \phi_e(0) + C_0^o Z^X\} .$$

Substituting $C_1^o = ZC_0^o$ in the second expression and collecting terms yields

$$C_0^o(X, 0, s) = \frac{\beta^{-1} F_0^o + \phi_e(0) + g(X)}{\beta^{-1}(1-\alpha Z) - Z^X}$$

$$= \frac{\{\beta^{-1} F_0^o + \phi_e(0)\} + g(X)}{Z^{-1} - Z^X} .$$

The last equality follows since Z satisfies $\alpha Z^2 - Z + \beta = 0$, by Theorem 2.3. Thus, $\phi(0,s) = \phi_a(0,s) + \phi_e(0)$, where $\phi_a(0,s) = \beta^{-1} F_0^o$, independent of $s \geq 0$. □

Denote

$$F^n = (F^n_{-(n-1)L+s-1}, F^n_{-(n-1)L+s-2}, \cdots, F^n_{-nL})^T \in \mathbb{R}^{s+L} .$$

Let $\{e(j)\}$ denote the unit vectors (where the dimension of $e(j)$ is determined by the context in which it is used); that is, $(e(j))_j = 1$ and $(e(j))_i = 0$, $i \neq j$, for $j = 0, 1, 2, \ldots$.

Theorem 2.21. Let $s \geq 1$. Denote $m = s+L-1$, $A = A(m)$, $\sigma(0,s) = 0$ and $\phi_a(0,s) = \beta^{-1} F_0^o$. Then, for $N \geq 1$,

$$\phi(N,s) = \phi_a(N,s) + \phi_e(N) \ , \tag{2.29}$$

where

$$\phi_a(N,s) = \frac{1}{\beta A^{-1}_{sm}} \{\phi_a(N-1,s) + A^{-1}_{s.}(F^N + \alpha\sigma(N-1,s)\ e(0))\} \tag{2.30}$$

and

$$\sigma(N,s) = A^{-1}_{L-1,.}\{F^N + \alpha\sigma(N-1,s)\ e(0) - \beta\phi_a(N,s)\ e(m)\} \ . \tag{2.31}$$

Proof. The proof is lengthy and is, therefore, only outlined here. See Theorem 3.21 of [9] for details.

The proof was the "dummy variable" substitution C_d previously discussed in Section 2.6. For $N = 1$, the transition equations form a Q-system of dimension $m + 2$. Recall (Section 2.6) that the first $N(s+L)$ equations of the linear systems for N and $N + 1$ are identical. In general, the proper substitution of the results for N into the transition equations for $N + 1$ will similarly yield a Q-system of dimension $m + 2$. The theorem then follows by induction on N using Lemma 2.23. □

Theorem 2.21 provides the necessary recursion. However, in order to implement the recursion, $A_{s.}^{-1}$ and $A_{L-1,.}^{-1}$ must first be computed (where $A = A(s+L-1)$. The next lemma provides a simple means for computing $A_{s.}^{-1}$ and $A_{L-1,.}^{-1}$.

Lemma 2.22. $A^{-1}(0) = [1]$ and for $m \geq 0$,

$$A_{i.}^{-1}(m+1) = (A_{i.}^{-1}(m) + \alpha\beta\delta(m)\, A_{im}^{-1}(m)\, A_{m.}^{-1}(m),\ \beta\delta(m)\, A_{im}^{-1}(m)),$$

$$i = 0, 1, \ldots, m \qquad (2.32)$$

and

$$A_{m+1,.}^{-1}(m+1) = (\alpha\delta(m)\, A_{m.}^{-1}(m),\ \delta(m)) , \qquad (2.33)$$

where

$$\delta(m) = (1 - \alpha\beta A_{mm}^{-1}(m))^{-1} . \qquad (2.34)$$

Proof. By Definition 2.5, $A(0) = [1] = A^{-1}(0)$; also,

$$A(m+1) = \begin{bmatrix} A(m) & B \\ C & 1 \end{bmatrix} , \qquad (2.35)$$

where

$$B = (0, 0, \ldots, 0, -\beta)^T \in \mathbb{R}^{m+1} \tag{2.35a}$$

$$C = (0, 0, \ldots, 0, -\alpha); \quad C^T \in \mathbb{R}^{m+1} \ . \tag{2.35b}$$

Partition $A^{-1}(m+1)$ as

$$A^{-1}(m+1) = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \delta(m) \end{bmatrix} \ , \tag{2.36}$$

where $\mathcal{A} \in \mathbb{R}^{(m+1)\times(m+1)}$, $\mathcal{B} \in \mathbb{R}^{m+1}$, $\mathcal{C}^T \in \mathbb{R}^{m+1}$, and $\delta(m) \in \mathbb{R}$. Since $A(m+1)\ A^{-1}(m+1) = I$, it follows from (2.35) and (2.36) that

$$A(m)\ \mathcal{A} + B\mathcal{C} = I \ , \tag{2.37a}$$

$$A(m)\ \mathcal{B} + \delta(m)B = (0, \ldots, 0)^T \ , \tag{2.37b}$$

$$C\mathcal{A} + \mathcal{C} = (0, \ldots, 0) \ , \tag{2.37c}$$

$$C\mathcal{B} + \delta(m) = 1 \ . \tag{2.37d}$$

From (2.37b),

$$\mathcal{B} = -\delta(m)\ A^{-1}(m)B \ . \tag{2.38}$$

Substituting (2.38) into (2.37d) gives

$$-\delta(m)\ CA^{-1}(m)B + \delta(m) = 1$$

$$\Rightarrow \delta(m) = (1 - CA^{-1}(m)B)^{-1} \quad . \tag{2.39}$$

Using (2.35a),

$$A^{-1}(m)B = -\beta A^{-1}_{.m}(m) \quad . \tag{2.40}$$

Hence, using (2.35b),

$$CA^{-1}(m)B = \alpha\beta A^{-1}_{mm}(m) \quad .$$

Substituting the above expression into (2.39) gives

$$\delta(m) = (1 - \alpha\beta A^{-1}_{mm}(m))^{-1} \ ,$$

which verifies (2.34).

Using (2.37a),

$$\alpha = A^{-1}(m) - A^{-1}(m)B\alpha \quad . \tag{2.41}$$

From (2.37c), (2.39) and (2.41),

$$C\mathcal{Q} + \mathcal{C} = C(A^{-1}(m) - A^{-1}(m)BC) + \mathcal{C}$$

$$= CA^{-1}(m) - (CA^{-1}(m)B-1)\mathcal{C} = 0$$

$$\Rightarrow$$

$$CA^{-1}(m) = (CA^{-1}(m)B - 1)\mathcal{C}$$

$$\Rightarrow$$

$$\mathcal{C} = (CA^{-1}(m)B - 1)^{-1} CA^{-1}(m)$$

$$= -\delta(m)\ CA^{-1}(m) \quad . \tag{2.42}$$

Using (2.35b),

$$CA^{-1}(m) = -\alpha A^{-1}_{m.}(m) \quad . \tag{2.43}$$

Substituting (2.43) into (2.42) yields

$$\mathcal{C} = \alpha\delta(m)\ A^{-1}_{m.}(m) \quad . \tag{2.44}$$

Substituting (2.40) into (2.38) similarly gives

$$\mathcal{B} = \beta\delta(m)\ A^{-1}_{.m}(m) \quad . \tag{2.45}$$

Using (2.40) and (2.44) yields

$$A^{-1}(m)B\mathcal{C} = -\beta A^{-1}_{\cdot m}(m)\ \delta(m)\alpha\ A^{-1}_{m\cdot}(m)$$

$$= -\alpha\beta\delta(m)\ A^{-1}_{\cdot m}(m)\ A^{-1}_{m\cdot}(m) \quad .$$

Substituting the above expression into (2.41) gives

$$\mathcal{A} = A^{-1}(m) + \alpha\beta\delta(m)\ A^{-1}_{\cdot m}(m)\ A^{-1}_{m\cdot}(m) \quad . \tag{2.46}$$

Using (2.36), (2.32) follows from (2.45) and (2.46); (2.33) follows similarly from (2.44). □

Using Lemma 2.22, a simple algorithm for computing $A^{-1}_{L-1,\cdot}(s+L-1)$ and $A^{-1}_{s\cdot}(s+L-1)$ can be written. Let $m_1 = \min\{s, L-1\}$ and $m_2 = \max\{s, L-1\}$. Then the algorithm can be outlined as follows:

(a) Compute $A^{-1}_{m_1\cdot}(m_1)$ using (2.33), for $m = 0, \ldots, m_1-1$.

(b) Compute $A^{-1}_{m_1\cdot}(m_2)$ using (2.32) and $A^{-1}_{m_2\cdot}(m_2)$ using (2.33), for $m = m_1, \ldots, m_2-1$.

(c) Compute $A^{-1}_{m_1\cdot}(s+L-1)$ and $A^{-1}_{m_2\cdot}(s+L-1)$ using (2.32), for $m = m_2, \ldots, s+L-2$.

It should be noted that (2.33) must also be used in part (c) in order to maintain $A^{-1}_{m\cdot}(m)$ for $m = m_2, \ldots, s+L-2$. The actual algorithm is given below and uses three vectors B^i, $i = 1, 2, 3$, of arbitrary dimension.

Algorithm B_1 (Compute $B^1 = A^{-1}_{s.}(s+L-1)$ and $B^2 = A^{-1}_{L-1,.}(s+L-1)$).

(1) If $s > L-1$, go to step (2); otherwise, $i \leftarrow 1$, $j \leftarrow 2$, $m_1 \leftarrow s$, $m_2 \leftarrow L-1$, $m \leftarrow 0$, $m_3 \leftarrow s+L-1$, $B^1 \leftarrow (1)$ and go to step (3).

(2) $i \leftarrow 2$, $j \leftarrow 1$, $m_1 \leftarrow L-1$, $m_2 \leftarrow L-1$, $m \leftarrow 0$, $m_3 \leftarrow s+L-1$, $B^2 \leftarrow (1)$.

(3) If $m \geq m_1-1$, go to step (5); otherwise, continue.

(4) $\delta \leftarrow (1 - \alpha\beta B^i_m)^{-1}$, $B^i \leftarrow (\alpha\delta B^i, \delta)$, $m \leftarrow m+1$; go to step (3).

(5) $B^j \leftarrow B^i$.

(6) If $m \geq m_2-1$, go to step (8); otherwise, continue.

(7) $\delta \leftarrow (1 - \alpha\beta B^j_m)^{-1}$, $B^i \leftarrow (B^i + \alpha\beta\delta B^i_m B^j, \beta\delta B^i_m)$, $B^j \leftarrow (\alpha\delta B^j, \delta)$, $m \leftarrow m+1$; go to step (6).

(8) $B^3 \leftarrow B^j$.

(9) If $m \geq m_3-1$, stop; otherwise, continue.

(10) $\delta \leftarrow (1 - \alpha\beta B^3_m)^{-1}$, $B^1 \leftarrow (B^1 + \alpha\beta\delta B^1_m B^3, \beta\delta B^1_m)$, $B^2 \leftarrow (B^2 + \alpha\beta\delta B^2_m B^3, \beta\delta B^2_m)$, $B^3 \leftarrow (\alpha\delta B^3, \delta)$, $m \leftarrow m+1$; go to step (9).

When the algorithm terminates in step (9), vector B^1 is $A^{-1}_{s.}(s+L-1)$ and vector B^2 is $A^{-1}_{L-1,.}(s+L-1)$. Given these vectors, the recursion of Theorem 2.21 can then be used to compute the coefficients $\phi(\cdot,s)$.

Algorithm B_2 (Compute $\phi(N,s)$, $N = 0, 1, 2, \ldots, \bar{N}$).

(1) Compute vectors B^1 and B^2 (use Algorithm B_1).

(2) $\phi_a \leftarrow \beta^{-1} F^o_0$, $\sigma \leftarrow 0$, $m \leftarrow s+L-1$, $N \leftarrow 0$, $a \leftarrow (\beta B^1_m)^{-1}$.

(3) $\phi(N,s) \leftarrow \phi_a + \phi_e(N)$.

(4) If $N \geq \bar{N}$, stop; otherwise, continue.

(5) $N \leftarrow N+1$, $B^3 \leftarrow F^N$, $B^3_0 \leftarrow B^3_0 + \alpha\sigma$, $\phi_a \leftarrow a(\phi_a + B^1 \cdot B^3)$, $B^3_m \leftarrow B^3_m - \beta\phi_a$, $\sigma \leftarrow B^2 \cdot B^3$; go to step (3).

The vector B^3 used in step (5) is a work vector. The calculations in step (5) follow the recursions (2.30) and (2.31), given $B^1 = A_{s.}^{-1}(s+L-1)$ and $B^2 = A_{L-1,.}^{-1}(s+L-1)$ from Algorithm B_1.

Algorithms B_1 and B_2 provide a means for determining the functionals $\{C_0^o(\cdot,N,s),\ N \geq 0\}$ for a fixed given value $s \geq 1$. By Theorem 2.11, only the functionals $\{C_0^o(\cdot,N,s^*),\ N \geq 0\}$ need be computed. Therefore, the next section focuses on the problem of determining the optimal disconnection parameter s^* prior to initiating Algorithms B_1 and B_2.

2.8. Determining s^*

In this section, a procedure is derived for computing s^*, the optimal disconnection parameter. By Lemma 2.3, the value s^* is independent of X, N, γ_1 and g. Therefore, let $X = L$, $N = 1$, $\gamma_1 = 0$, $g \equiv 0$ and denote

$$\bar{C}_0^o(s) = C_0^o(L,1,s;\ g \equiv 0,\ \gamma_1 = 0), \qquad s \geq 0$$

The only costs contributing to $\bar{C}_0^o(s)$ are the charges c, d, π and $\phi_e(1)$. $\phi_e(1)$ is incurred only at expansion epochs and those epochs are independent of the disconnect parameter s. Therefore, only the temporary module costs resulting from connection (c), disconnection (d), and rental (π) charges vary with s. That is, the difference $\bar{C}_0^o(s+1) - \bar{C}_0^o(s)$ must be a positive affine multiple of $-H(s+1,s)$, $s \geq 0$. Hence it follows from Definition 2.4 that

$$\bar{C}_0^o(s^*) = \min_{s \geq 0} \bar{C}_0^o(s) \quad . \tag{2.47}$$

Furthermore, it follows from Theorem 2.8 that $-\bar{C}_0^o(\cdot)$ is strictly unimodal about $[s^*-1, s^*]$.

The results of the last section provide expressions for $\bar{C}_0^o(s)$, $s \geq 1$. When $s = 0$, Model II reduces to Model I. The next lemma provides a means for identifying instances where $s^* = 0$; in these cases, Model I will be applicable. Proceeding as before, the transition equations for the case $s = 0$ and $N = 1$ are

$$C_0^o(X,1,0) - \beta(1-\alpha Z)^{-1} C_{-1}^o(X,1,0) = (1-\alpha Z)^{-1} F_0^o ,$$

$$-\alpha C_{k+1}^1(X,1,0) + C_k^1(X,1,0) - \beta C_{k-1}^1(X,1,0) = F_k^1, \qquad -L+1 \leq k \leq -1$$

$$-\alpha C_{-L+1}^1(X,1,0) + C_{-L}^1(X,1,0) = F_{-L}^1 + \beta\{\phi_e(1) + g(X) + Z^{X-L} C_0^o(X,1,0)\} .$$

Lemma 2.23. Denote $A = A(L-1)$. Then

$$C_0^o(X,1,0) = \frac{\phi_a(1,0) + \phi_e(1) + g(X)}{Z^{-1} - Z^X} Z^L , \tag{2.48}$$

where

$$\phi_a(1,0) = \frac{1}{\beta A_{0,L-1}^{-1}} (\beta^{-1} F_0^o + A_{0.}^{-1} F^1) \quad . \tag{2.49}$$

Proof. See Lemma 3.23 of [9].

Denote $\bar{\phi}_a(s) = \phi_a(1,s;\ r_1 = 0)$, $s \geq 0$. Then, by Theorem 2.21 and Lemma 2.23,

$$\bar{C}_0^o(s) = \frac{\bar{\phi}_a(s) + \phi_e(1)}{z^{-1} - z^L} z^L \quad , \qquad s \geq 0 \ . \tag{2.50}$$

Hence, by (2.47)

$$\bar{\phi}_a(s^*) = \min_{s \geq 0} \bar{\phi}_a(s) \ , \tag{2.51}$$

and furthermore, $-\bar{\phi}_a(\cdot)$ is unimodal about $[s^*-1, s^*]$. Using Theorem 2.21 and Lemma 2.23, a recursive procedure will be derived for evaluating $\{\bar{\phi}_a(s),\ s = 0,\ 1,\ \ldots\}$ in order to determine s^*.

Lemma 2.24. $A^{-1}(0) = [1]$ and for $m \geq 0$,

$$A_{0.}^{-1}(m+1) = (\delta'(m),\ \beta\delta'(m)\ A_{0.}^{-1}(m)) \ , \tag{2.52}$$

and

$$A_{i.}^{-1}(m+1) = (\alpha\delta'(m)\ A_{i-1,0}^{-1}(m),\ A_{i-1,.}^{-1}(m) + \alpha\beta\delta'(m)\ A_{i-1,0}^{-1}(m)\ A_{0.}^{-1}(m)),$$

$$i = 1,\ \ldots,\ m+1 \tag{2.53}$$

where

$$\delta'(m) = (1 - \alpha\beta A_{00}^{-1}(m))^{-1} \ . \tag{2.54}$$

Proof. By Definition 2.5, $A(0) = [1] = A^{-1}(0)$; also,

$$A(m+1) = \begin{bmatrix} 1 & B \\ C & A(m) \end{bmatrix} , \tag{2.55}$$

where $B = (-\beta, 0, \ldots, 0)$; $B^T \in \mathbb{R}^{m+1}$, and $C = (-\alpha, 0, \ldots, 0)^T \in \mathbb{R}^{m+1}$. Partition $A^{-1}(m+1)$ in a manner analogous to (2.55). The lemma then follows as in the proof of Lemma 2.22 using $A^{-1}(m+1)\, A(m+1) = I$. See Lemma 3.24 of [9] for details. □

Definition 2.7. For $s \geq 0$, let

$$a_1(s) = A_{00}^{-1}(s+L-1) ,$$

$$a_2(s) = A_{s0}^{-1}(s+L-1) ,$$

$$a_3(s) = A_{0,s+L-1}^{-1}(s+L-1) ,$$

$$a_4(s) = A_{s,s+L-1}^{-1}(s+L-1) ,$$

$$\mathcal{E}_1(s) = A_{s.}^{-1}(s+L-1)\, \underset{\sim}{1} ,$$

$$\mathcal{E}_2(s) = A_{0.}^{-1}(s+L-1)\, \underset{\sim}{1} ,$$

where

$$\underset{\sim}{1} = (1, 1, \ldots, 1)^T .$$

Lemma 2.25. For $m \geq 0$,

$$A_{00}^{-1}(m+1) = \delta'(m) , \tag{2.56}$$

$$A_{0,m+1}^{-1}(m+1) = \beta\delta'(m)\, A_{0m}^{-1}(m) , \tag{2.57}$$

and

$$A_{0\cdot}^{-1}(m+1)\underset{\sim}{1} = \delta'(m)\{1 + \beta A_{0\cdot}^{-1}(m)\underset{\sim}{1}\} \quad , \tag{2.58}$$

where $\delta'(m)$ is given by (2.54).

Proof. The results follow directly from (2.52) and (2.54) of the previous lemma. □

Lemma 2.26. For $s \geq 0$,

$$\begin{aligned}
a_1(s+1) &= (1 - \alpha\beta a_1(s))^{-1} \quad , \\
a_2(s+1) &= \alpha a_1(s+1)\, a_2(s) \quad , \\
a_3(s+1) &= \beta a_1(s+1)\, a_3(s) \quad , \\
a_4(s+1) &= a_4(s) + \beta a_2(s+1)\, a_3(s) \\
\varepsilon_1(s+1) &= \varepsilon_1(s) + a_2(s+1)\{1 + \beta\varepsilon_2(s)\} \\
\varepsilon_2(s+1) &= a_1(s+1)\{1 + \beta\varepsilon_2(s)\} \quad .
\end{aligned} \tag{2.59}$$

Proof. Using Definition 2.7 with Lemmas 2.24 and 2.25,

$$\begin{aligned}
a_1(s+1) = A_{00}^{-1}(s+L) = \delta'(s+L-1) &= (1 - \alpha\beta A_{00}^{-1}(s+L-1))^{-1} \\
&= (1 - \alpha\beta a_1(s))^{-1}
\end{aligned}$$

$$\begin{aligned}
a_2(s+1) = A_{s+1,0}^{-1}(s+L) &= \alpha\delta'(s+L-1)\, A_{s0}^{-1}(s+L-1) \\
&= \alpha a_1(s+1)\, a_2(s)
\end{aligned}$$

$$a_3(s+1) = A^{-1}_{0,s+L}(s+L) = \beta\delta'(s+L-1)\ A^{-1}_{0,s+L-1}(s+L-1)$$

$$= \beta a_1(s+1)\ a_3(s)$$

$$a_4(s+1) = A^{-1}_{s+1,s+L}(s+L)$$

$$= A^{-1}_{s,s+L-1}(s+L-1) + \alpha\beta\delta'(s+L-1)\ A^{-1}_{s0}(s+L-1)\ A^{-1}_{0,s+L-1}(s+L-1)$$

$$= a_4(s) + a_2(s+1)\ \beta A^{-1}_{0,s+L-1}(s+L-1)$$

$$= a_4(s) + \beta a_2(s+1)\ a_3(s)$$

$$\varepsilon_1(s+1) = A^{-1}_{s+1,\cdot}(s+L)\ \underset{\sim}{1}$$

$$= \alpha\delta'(s+L-1)\ A^{-1}_{s0}(s+L-1) + A^{-1}_{s\cdot}(s+L-1)\ \underset{\sim}{1}$$

$$+ \alpha\beta\delta'(s+L-1)\ A^{-1}_{s0}(s+L-1)\ A^{-1}_{0\cdot}(s+L-1)\underset{\sim}{1}$$

$$= \varepsilon_1(s) + \alpha\delta'(s+L-1)\ A^{-1}_{s0}(s+L-1)\ \{1 + \beta\varepsilon_2(s)\}$$

$$= \varepsilon_1(s) + a_2(s+1)\ \{1 + \beta\varepsilon_2(s)\}$$

$$\varepsilon_2(s+1) = A^{-1}_{0\cdot}(s+L)\underset{\sim}{1}$$

$$= \delta'(s+L-1) + \beta\delta'(s+L-1)\ A^{-1}_{0\cdot}(s+L-1)\underset{\sim}{1}$$

$$= \delta'(s+L-1)\ \{1 + \beta A^{-1}_{0\cdot}(s+L-1)\underset{\sim}{1}\}$$

$$= a_1(s+1)\ \{1 + \beta\varepsilon_2(s)\}\ .$$ □

The significance of Definition 2.7 is given by the following theorem.

Theorem 2.27.

$$\bar{\phi}_a(s) = c + (\beta a_4(s))^{-1} \{c + \alpha d\, a_2(s) + \Pi \varepsilon_1(s)\}, \quad s \geq 0 \ . \tag{2.60}$$

Proof. By Theorem 2.21 and Lemma 2.23,

$$\bar{\phi}_a(s) = \frac{1}{\beta A^{-1}_{s,s+L-1}(s+L-1)} \{\beta^{-1} F^o_0 + A^{-1}_{s\cdot}(s+L-1)\, F^1\} \ . \tag{2.61}$$

Given that $\gamma_1 = 0$,

$$F^o_0 = \beta c,$$

and

$$F^1 = (\alpha d + \Pi, \Pi, \Pi, \ldots, \Pi, \beta c + \Pi) \in \mathbb{R}^{s+L} \ .$$

Substituting these expressions into (2.61) and Definition 2.7 yields

$$\bar{\phi}_a(s) = \frac{1}{\beta a_4(s)} \{c + \alpha d A^{-1}_{s0}(s+L-1)\, \beta c A^{-1}_{s,s+L-1}(s+L-1) + \Pi A^{-1}_{s\cdot}(s+L-1)\underset{\sim}{1}\}$$

$$= \frac{1}{\beta a_4(s)} \{c + \alpha d a_2(s) + \beta c a_4(s) + \Pi \varepsilon_1(s)\}$$

$$= c + (\beta a_4(s))^{-1} \{c + \alpha d a_2(s) + \Pi \varepsilon_1(s)\} \ . \qquad \square$$

The desired recursion can now be outlined as follows:

(a) Use (2.56) - (2.58) to compute

$$a_1(0) = a_2(0) = A_{00}^{-1}(L\text{-}1) ,$$

$$a_3(0) = a_4(0) = A_{0,L-1}^{-1}(L\text{-}1) ,$$

and

$$\varepsilon_1(0) = \varepsilon_2(0) = A_{0,\cdot}^{-1}(L\text{-}1)\underset{\sim}{1} .$$

(b) Recursively compute $\bar{\phi}_a(s)$, $s \geq 0$ using (2.60) with (2.59) until s^* is identified by $\bar{\phi}_a(s^*) < \bar{\phi}_a(s^*+1)$.

The relationship identifying s^* in part (b) follows from the unimodal property of $-\bar{\phi}_a(\cdot)$. The algorithm is stated below.

Algorithm B_3 (Find s^*).

(1) $a_1 \leftarrow 1$, $a_3 \leftarrow 1$, $\varepsilon_1 \leftarrow 1$, $m \leftarrow 0$.

(2) If $m \geq L\text{-}1$, go to step (4); otherwise, continue.

(3) $a_1 \leftarrow (1 - \alpha\beta a_1)^{-1}$, $a_3 \leftarrow \beta a_1 a_3$, $\varepsilon_1 \leftarrow a_1(1 + \beta\varepsilon_1)$, $m \leftarrow m+1$; go to step (2).

(4) $s^* \leftarrow 0$, $a_2 \leftarrow a$, $a_4 \leftarrow a_3$, $\varepsilon_2 \leftarrow \varepsilon_1$, $\hat{\phi}_a \leftarrow (\beta a_4)^{-1}(c + \alpha d a_2 + \pi\varepsilon_1)$.

(5) $a_1 \leftarrow (1\text{-}\alpha\beta a_1)^{-1}$, $a_2 \leftarrow \alpha a_1 a_2$, $a_4 \leftarrow a_4 + \beta a_2 a_3$, $a_3 \leftarrow \beta a_1 a_3$, $v \leftarrow 1+\beta\varepsilon_2$, $\varepsilon_1 \leftarrow \varepsilon_1 + a_2 v$, $\varepsilon_2 \leftarrow a_1 v$, $\bar{\phi}_a \leftarrow (\beta a_4)^{-1}(c + \alpha d a_2 + \pi\varepsilon_1)$.

(6) If $\bar{\phi}_a > \hat{\phi}_a$, stop; otherwise, $\hat{\phi}_a \leftarrow \bar{\phi}_a$, $s^* \leftarrow s^*+1$, and go to step (5).

Steps (1) - (3) of the algorithm compute $a_1(0)$, $a_3(0)$ and $\varepsilon_1(0)$ using (2.56) - (2.58). Step (4) computes $\bar{\phi}_a(0)$ - c. Step (5) successively computes $\bar{\phi}_a(s)$ - c, $s \geq 1$, and step (6) compares $\bar{\phi}_a(s)$ - c with

$\tilde{\phi}_a(s-1) - c$. The algorithm terminates since $s^* < +\infty$, by Lemma 2.2. When termination occurs, the optimal disconnection parameter is given by s^*.

2.9. Summary and Statement of the Expansion Size Optimization Problem

Let $\kappa = \{0, L, 2L, \ldots, \tilde{N}L\}$, where $\tilde{N}$ denotes an upper bound on temporary module usage based on physical considerations. For $K \in \kappa$, denote

$$\phi(K) = \phi(K/L, s^*) \quad \text{and} \quad C_k(X,K) = C_k^o(X, K/L, s^*), \quad k \geq 0 . \tag{2.62}$$

Given $K \in \kappa$ and an initial spares level $k_0 \geq 0$, an optimal permanent expansion size $X^*(K)+1$ for Model II is given by

$$C_0(X^*(K), K) = \min\{C_0(X,K) : \text{integer } X \geq K\} , \tag{2.63}$$

where

$$C_0(X,K) = \frac{\phi(K) + g(X)}{z^{-1} - z^X} z^K , \qquad X \geq K . \tag{2.64}$$

$\phi(\cdot)$ is nonnegative and nondecreasing over κ; (2.65)

$0 < z = \frac{1 - \sqrt{1-4\alpha\beta}}{2\alpha} - 1$, given α and β satisfying (2.13) $(\alpha z^2 - z + \beta = 0)$; (2.66)

$$C_k(X,K) = Z^k\, C_0(X,K), \qquad k \geq 0\ ; \tag{2.67}$$

and

$$Z^k = E[e^{-rT_k}], \qquad k \geq 1\ . \tag{2.68}$$

(2.64) follows from Theorem 2.17, (2.65) follows from Theorem 2.17 and Lemma 2.18, (2.66) and (2.67) follow from Theorem 2.13, and (2.68) follows from Lemma 2.15. The fact that $X^*(K)$ is given by (2.63), independent of $k_0 \geq 0$, follows from Corollary 2.14.

Algorithm B_3, presented in the previous section, provides a simple means for determining s^*. If $s^* = 0$, then Model I of [7] is applicable. If $s^* \neq 0$, then Algorithms B_1 and B_2 of Section 2.7 provide recursions for computing the coefficients $\phi(K) = \phi(K/L, s^*)$ over $K \in \kappa$. Once these coefficients are known, the task of determining $X^*(.)$ over κ becomes the series of single-variable minimization problems given by (2.63) under conditions (2.64) - (2.68). Notice that this minimization problem is identical to that resulting for Model I of [7]. The optimization problem is treated in [8]. By Corollary 2.14, the optimal temporary facilities usage limit K^* (alternatively, the optimal temporary modules limit $N^* = K^*/L$) and the associated optimal expansion size X^*+1 are then given by

$$C_0(X^*, K^*) = \min\{C_0(X^*(K), K) : K \in \kappa\}\ .$$

REFERENCES

[1] William Feller, An Introduction to Probability Theory and Its Applications, Vol. 1, Third Edition, John Wiley and Sons, New York, 1968.

[2] John Freidenfelds, "Cable Sizing with Stochastic Demand", Proceedings of the Sixth Annual Pittsburgh Conference on Modeling and Simulation, Instrument Society of America (pub.), April 1975.

[3] Francis B. Hildebrand, Finite-Difference Equations and Simulations, Prentice-Hall, 1968.

[4] Warren L.G. Koontz and R.S. Shipley, "Application of Subscriber Pair Gain Systems in an Environment of Stochastic Demand", Proceedings of the Sixth Annual Pittsburgh Conference on Modeling and Simulation, Instrument Society of America (pub.), April 1975.

[5] Alan S. Manne, "Capacity Expansion and Probabilistic Growth", Econometrica, Vol. 29, October 1961, pp. 632-649.

[6] Sheldon M. Ross, Applied Probability Models with Optimization Applications, Holden-Day, San Francisco, 1970.

[7] R. Scott Shipley, "A Stochastic Capacity Expansion Model: Non-Modular Temporary Facilities", Stanford University, Department of Operations Research, Technical Report No. 178, September 1976.

[8] R. Scott Shipley, "Optimization of Recurrent Stochastic Capacity Expansion Models and Generalization to a Non-Recurrent Model", Stanford University, Department of Operations Research, Technical Report No. 180, October 1976.

[9] Robert Scott Shipley, " Stochastic Capacity Expansion Models", Ph.D. Dissertation, Stanford University, September 1976.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER (14) TR-179	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) A Stochastic Capacity Expansion Model: Modular Temporary Facilities,	5. TYPE OF REPORT & PERIOD COVERED (9) Technical Report, / 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) R. Scott Shipley	8. CONTRACT OR GRANT NUMBER(s) (15) N00014-75-C-0561
9. PERFORMING ORGANIZATION NAME AND ADDRESS Dept. of Operations Research & Dept. of Statistics, Stanford University, Stanford, California 94305	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS (NR-042-002)
11. CONTROLLING OFFICE NAME AND ADDRESS Statistics & Probability Program Code 436 Office of Naval Research Arlington, Virginia 22217	12. REPORT DATE (11) 4 Oct[redacted]976 / 13. NUMBER OF PAGES 80
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) (12) 85 p.	15. SECURITY CLASS. (of this report) Unclassified / 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION IS UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

BACKLOGGING, CAPACITY EXPANSION, INVENTORY THEORY, JOBLETTING, OVERLOADING. POISSON PROCESSES

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See reverse side.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

4/02 766 Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

mt

20. Abstract

This paper considers optimal decision strategies with regard to capacity expansion in an environment where demand arrivals and departures can be characterized as independent Poisson processes. Two types of facilities are considered in the model: permanent and temporary. Permanent facilities represent the means by which demand is normally served, while temporary facilities represent the extraordinary measures taken in order to serve excess demand (prior to an expansion of permanent facilities). Examples of temporary facilities include backlogging, overloading and jobletting. ~~A companion paper [7] treats the case of non-modular temporary facilities.~~ This paper considers modular temporary facilities which typically are available in a fixed increment size (a module) and incur instantaneous installation and removal charges in addition to normal usage costs. Because of these additional charges, a sub-optimization problem arises with regard to how modules should be used so as to properly balance the instantaneous charges with the usage costs in order to minimize overall expected costs. It is shown that an optimal module removal policy has the form: "remove a model only if at least s^* additional units of unused capacity will remain available". A simple algorithm is given for determining the optimal parameter s^*.

For a given limit (K) on temporary facility usage, the form of the expected discounted cost functional, parameterized in the expansion size (X+1), is derived. Recursions are given for determining these functionals over all feasible values for K. A sequel paper [8] treats the problem of minimizing the functionals in order to find optimal expansion sizes.